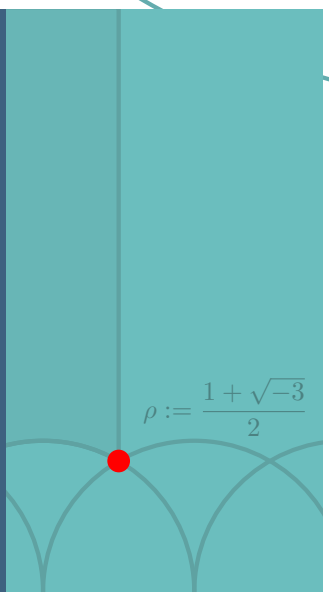
01/07/2026

Econometrics

Intermediate level notes



这是一份期末不完全总结笔记。

南开大学经济学院硕士生的中级计量经济学，主要使用的是林文夫（Hayashi）一书，这本书已经达到了高级计量的难度，在一些学校被用作三高教材。这本书的权威性勿庸置疑，好就好在建立了统一的GMM分析框架，但由于成书较早，我认为存在如下问题：

一是符号与现代计量教材不匹配，例如解释变量和工具变量的表示相反、前定变量的定义、各章节系数的表示、渐近方差的夹心三明治形式……
二是缺少完整严谨的证明，通常只给出证明的思路——这也是这份笔记尝试补充的；
三是缺少图形化解释。本笔记补充了投影矩阵的图形分解，Wald、LR、LM统计量三者的图象理解等。

这就造成了课程讲义同样存在上述问题，同时讲义存在严重的跳步（虽然这部分会由老师上课板书给出），但光看讲义就会云里雾里、不知所以然，回到书上一看又会豁然开朗，这种现象在使用Verbeek一书内容时尤为明显。

在学习中级计量的过程中，我主要参考了如下书籍：

- (荷)扬·R.麦纳斯著.计量经济学理论导论.格致出版社.2025
- 陈强编著.高级计量经济学及Stata应用.第2版.高等教育出版社.2014
- 洪永淼著.高级计量经济学.高等教育出版社.2011
- (美) 哈尔伯特·怀特著.计量经济学渐近理论.格致出版社.2023
- 威廉·H·格林著.计量经济分析.中国人民大学出版社.2020

对于初学者，我推荐第一本书作为前置阅读学习材料，这本书语言幽默风趣配有漫画，聚焦线性回归（含渐近理论）与极大似然。我在小红书上指出了中文版前两章的翻译错误，但由于未能找到完整的英文原版，对于整本书的纠错暂且搁置。

陈强老师在他书的前言部分写到**“其中尤以Hayashi(2000)对本书的影响最深”**，陈书修改了解释变量与工具变量的写法，大体遵循林文夫的思路，对于命题的证明更加完整，我参考较多有很大启发。

洪永淼老师的书最新英文版有部分更新，但中文版已经足够，强烈推荐洪教主的高计网课，B站上有三个版本，这三个版本互为补充，可以根据学习内容与自身英文水平选择观看：

【厦门大学】高级计量经济学（洪永淼）_哔哩哔哩_bilibili
高级计量经济学（洪永淼：英文字幕2020版）_哔哩哔哩_bilibili
高级计量经济学II-洪永淼_哔哩哔哩_bilibili

有学长评价（知乎@梅林的胡子i）：

> 推荐这本书倒不是他写得有多好，而是它的配套网课就能在b站找到，并且这可能是你能在全网找到的讲得最好的计量经济学网课。
> 幸好后来b站刷到了洪永淼老师的高级计量课程，听了之后简直是开了我的天眼了，洪教主讲课是真的细致，一步一步地给你推导每一个定理，证明该有的步骤是一个都不舍得跳过，一些重要的概念还会强调很多遍，非常适合我这种基础差的，我觉得只要上过概率论与数理统计，哪怕是本科生，学洪永淼的网课也没有一点问题。

如果看findingnothing的视频需要用到的书目为Greene的《Econometrics Analysis》和Hansen的《Econometrics》：findingnothing的个人空间-合集·计量经济学-哔哩哔哩视频

本系列笔记构成如下：第一章基于洪书第三章，基础内容完全够用；第二章忽略了4种不同的收敛形式与大数定律与中心极限定理的具体内容，因为在之后的证明中可以直接使用；第三章忽略了对极大似然估计原理的理解解释以及不考的LM检验；第四章由于2SLS的证明被纳入第五章GMM，不在考试重点所以较为简略，这部分的详细内容同样可以参考洪书第七章；第五章根据考试要求对不同Propositions有所选取。最后三章多方程GMM、受限因变量模型、面板GMM主要按照考点进行梳理，并根据教材进行了必要的补充。

随笔记附上一套金融学院中级计量经济学 Intermediate Econometrics | (Albert) Bo ZHAO 赵博期末试题，虽然题型不同，但对于加深各部分理解大有裨益。

感谢王子铭通读本笔记帮助修改细节错误，感谢王昊阳分享对计量精彩深入的理解。

本笔记已尽可能统一符号，并更正了一些讲义上的错误，但错误总是在所难免的，还望多多海涵指正，也欢迎与我联系交流。

我的个人网站：xishanyu2.github.io，笔记如有更正将会在上面更新。我的邮箱：zzynankai@outlook.com

西山yu
2025年12月29日

This is a note from Yongmiao Hong's book, I've changed the subscript t into i, this note covers the entire content of chap1.

Assumptions

Ass 1: Linearity

$$Y_i = X_i'\beta + \varepsilon_i$$

Ass 2: Strict Exogeneity

$$E(\varepsilon_i|X_i) = 0$$

By the law of iterated expectations(IIE):
$E(X_i\varepsilon_i) = 0$
$E(\varepsilon_i) = 0$
then $\mathrm{Cov}(X_i, \varepsilon_i) = 0$

Ass 3: Nonsingularity

$$Rank(\mathbf{X}'\mathbf{X}) = K \leq n$$

Ass 4: Spherical Error Variance

$$E(\varepsilon_i^2|\mathbf{X}) = \sigma^2$$

We can write Ass 2 and Ass 4 compactly as follows:

$$E(\varepsilon|\mathbf{X}) = 0 \text{ and } E(\varepsilon\varepsilon'|\mathbf{X}) = \sigma^2\mathbf{I}$$

Ass 5: Conditional Normality

$$\varepsilon|\mathbf{X} \sim N(0, \sigma^2\mathbf{I})$$

Ass 5 implies both Ass 2 and Ass 4.

OLS Estimator

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y \\ &= \left(\sum_{i=1}^n X_iX_i'\right)^{-1}\sum_{i=1}^n X_iY_i \\ &= \left(\frac{1}{n}\sum_{i=1}^n X_iX_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^n X_iY_i\end{aligned}$$

The FOC implies that the estimated residual e is **orthogonal** to regressors X:

$$\mathbf{X}'e = \sum_{i=1}^{n} X_i e_i = 0$$

Sample error:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$$

Define an $n \times n$ projection matrix:

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

and

$$\mathbf{M} = \mathbf{I} - \mathbf{P}$$

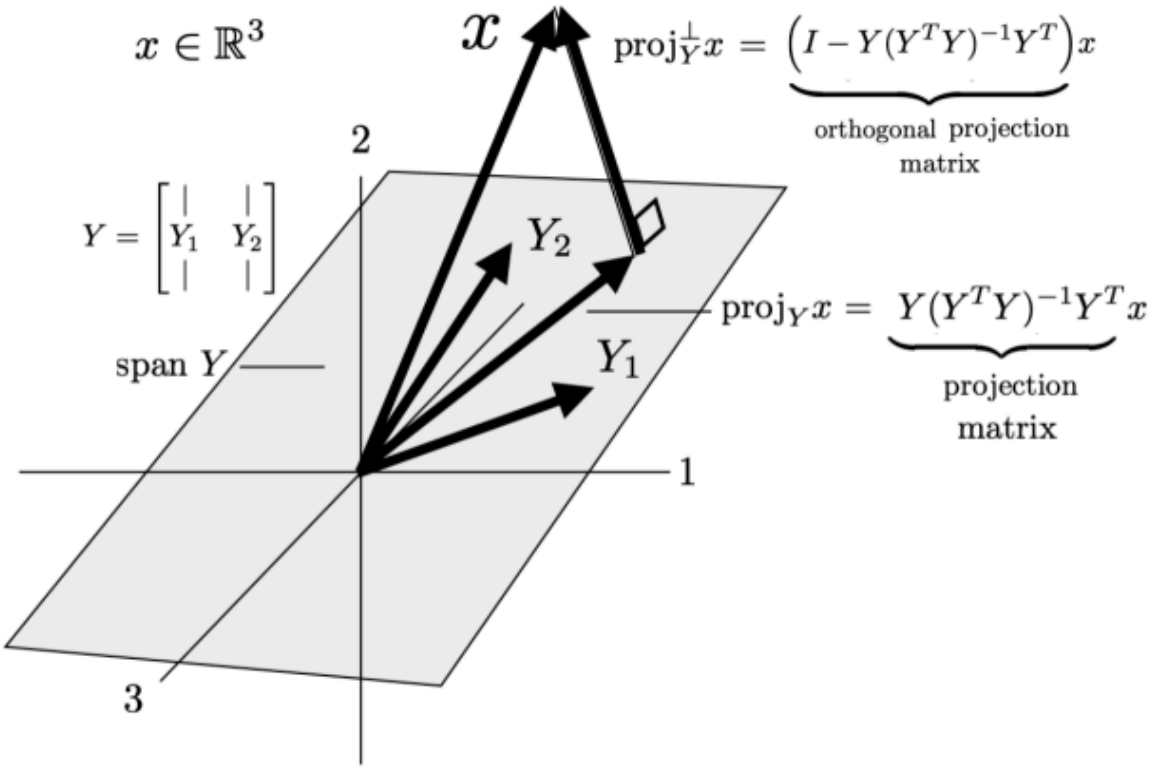
Then both matrices $\mathbf{P}$ and $\mathbf{M}$ are symmetric (i.e., $\mathbf{P} = \mathbf{P}'$ and $\mathbf{M} = \mathbf{M}'$) and idempotent (i.e., $\mathbf{P}^2 = \mathbf{P}, \mathbf{M}^2 = \mathbf{M}$), with

$$\mathbf{PX} = \mathbf{X},$$
$$\mathbf{MX} = 0$$

then:

$$SSR(\hat{\beta}) = e'e = Y'MY = \varepsilon'M\varepsilon$$

understanding from the projecting perspective:



proof: $P = \underset{n\times k}{X}\underset{k\times k}{(X'X)^{-1}}\underset{k\times n}{X'}$ $\quad M = I_n - P$

① $P' = P$:
$$\begin{aligned} P' &= (X(X'X)^{-1}X')' \\ &= X((X'X)^{-1})'X' \\ &= X((X'X)')^{-1}X' \\ &= X(X'X)^{-1}X' \\ &= P \end{aligned}$$

$M' = M$:
$$\begin{aligned} M' &= (I_n - P)' \\ &= I_n' - P' \\ &= I_n - P \\ &= M \end{aligned}$$

② $P^2 = P$:
$$\begin{aligned} P^2 &= X(X'X)^{-1}\cancel{X'X(X'X)^{-1}}X' \\ &= X(X'X)^{-1}X' \\ &= P \end{aligned}$$

$M^2 = M$:
$$\begin{aligned} M^2 &= (I_n - P)^2 \\ &= I_n^2 - I_nP - PI_n + P^2 \\ &= I_n - 2P + P \\ &= I_n - P \\ &= M \end{aligned}$$

③ $PX = X$: ↓ projection
$$\begin{aligned} PX &= X\cancel{(X'X)^{-1}X'X} \\ &= X \end{aligned}$$

$MX = 0$:
$$\begin{aligned} MX &= (I_n - P)X \\ &= X - PX \\ &= 0 \end{aligned}$$

④ $e = M\varepsilon$:
$$\begin{aligned} e &= Y - X\hat{\beta} \\ &= Y - X(X'X)^{-1}X'Y \\ &= (I - X(X'X)^{-1}X)Y \\ &= MY \\ &= M(X\beta^{\circ} + \varepsilon) \\ &= M\varepsilon \end{aligned}$$

⑤ $SSR(\hat{\beta}) = e'e$
$$\begin{aligned} &= (M\varepsilon)'M\varepsilon \\ &= \varepsilon'M'M\varepsilon \\ &= \varepsilon'MM\varepsilon \\ &= \varepsilon'M\varepsilon \end{aligned}$$

Goodness of Fit

Uncentered R^2:

$$R^2_{uc} = \frac{\hat{Y}'\hat{Y}}{Y'Y} = 1 - \frac{e'e}{Y'Y}$$

Centered R^2/Coefficient of Determination:

$$R^2 \equiv 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

$$R^2 = \rho^2_{Y\hat{Y}}, 0 \leq R^2 \leq 1$$

If the regression function doesn't contain the intercept, then the R^2 may be negative.

Consistency and Efficiency of the OLS Estimator

- Unbiasedness:
  $E(\hat{\beta}|\mathbf{X}) = \beta$ and $E(\hat{\beta}) = \beta$
- Vanishing Variance:
  var($\hat{\beta}|\mathbf{X}$) = $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$
- Orthogonality:
  cov($\hat{\beta}, e|\mathbf{X}$) = 0
- Gauss-Markov Theorem:
  var($\hat{b}|\mathbf{X}$) - var($\hat{\beta}|\mathbf{X}$) ~ PSD(positive semi-definite)
- Sample Residual Variance Estimator:

$$s^2 = e'e/(n-K) = \frac{1}{n-K}\sum_{i=1}^n e_i^2$$

Sampling Distribution of the OLS Estimator

Ass 5: Conditional Normality

$$\varepsilon|\mathbf{X} \sim N(0, \sigma^2\mathbf{I})$$

Normality of the OLS Estimator:

$$(\hat{\beta} - \beta)|\mathbf{X} \sim N[0, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]$$
$$R(\hat{\beta} - \beta)|\mathbf{X} \sim N[0, \sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R']$$

The $J \times K$ matrix R is a selection matrix.

Variance Estimation for the OLS Estimator

Residual Variance Estimator:

$$\left.\frac{(n-K)s^2}{\sigma^2}\right|_{\mathbf{X}} = \left.\frac{e'e}{\sigma^2}\right|_{\mathbf{X}} \sim \chi^2_{n-K}$$

Hypothesis Testing

$$\mathbf{H}_0 : R\beta = r$$

we can consider the statistic $R\hat{\beta} - r$ and check if this difference is significantly different from zero. Under H_0, we have:

$$R\hat{\beta} - r = R\hat{\beta} - R\beta = R(\hat{\beta} - \beta) \to 0 \text{ as } n \to \infty$$

Under the alternative to $\mathbf{H}_0$: $R\beta \neq r$, but we still have $\hat{\beta} - \beta \to 0$. It follows that:

$$R\hat{\beta} - r = R(\hat{\beta} - \beta) + R\beta - r \to R\beta - r \neq 0 \text{ as } n \to \infty$$

In other words, $R\hat{\beta} - r$ will converge to a nonzero limit $R\beta - r$.

Under H_0:

$$(R\hat{\boldsymbol{\beta}} - r)|\mathbf{X} \sim N(0, \sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R')$$

Questions:

(1) $\text{var}[(R\hat{\beta} - r)|\mathbf{X}] = \sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R'$ is a scalar or a matrix?

(2) Since σ^2 is unknown, $\frac{R\hat{\beta}-r}{\sqrt{\sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R'}}$ is no longer normally distributed.

Case 1: t-Test: $J = 1$, replace σ^2 by s^2

$$\begin{aligned}
\mathrm{T} &= \frac{R\hat{\beta} - r}{\sqrt{s^2 R(\mathbf{X}'\mathbf{X})^{-1}R'}} \\
&= \frac{\frac{R\hat{\beta}-r}{\sqrt{\sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R'}}}{\sqrt{\frac{(n-K)s^2}{\sigma^2}/(n-K)}} \\
&\sim \frac{N(0,1)}{\sqrt{\chi^2_{n-K}/(n-K)}} \\
&\sim t_{n-K}
\end{aligned}$$

- Decision Rule of the t-Test Based on Critical Values:

Reject H_0: $|T| > C_{t_{n-K}, \frac{\alpha}{2}}$

$$P\left(t_{n-K} > C_{t_{n-K},\frac{\alpha}{2}}\right) = \frac{\alpha}{2}$$
$$P\left(|t_{n-K}| > C_{t_{n-K},\frac{\alpha}{2}}\right) = \alpha$$

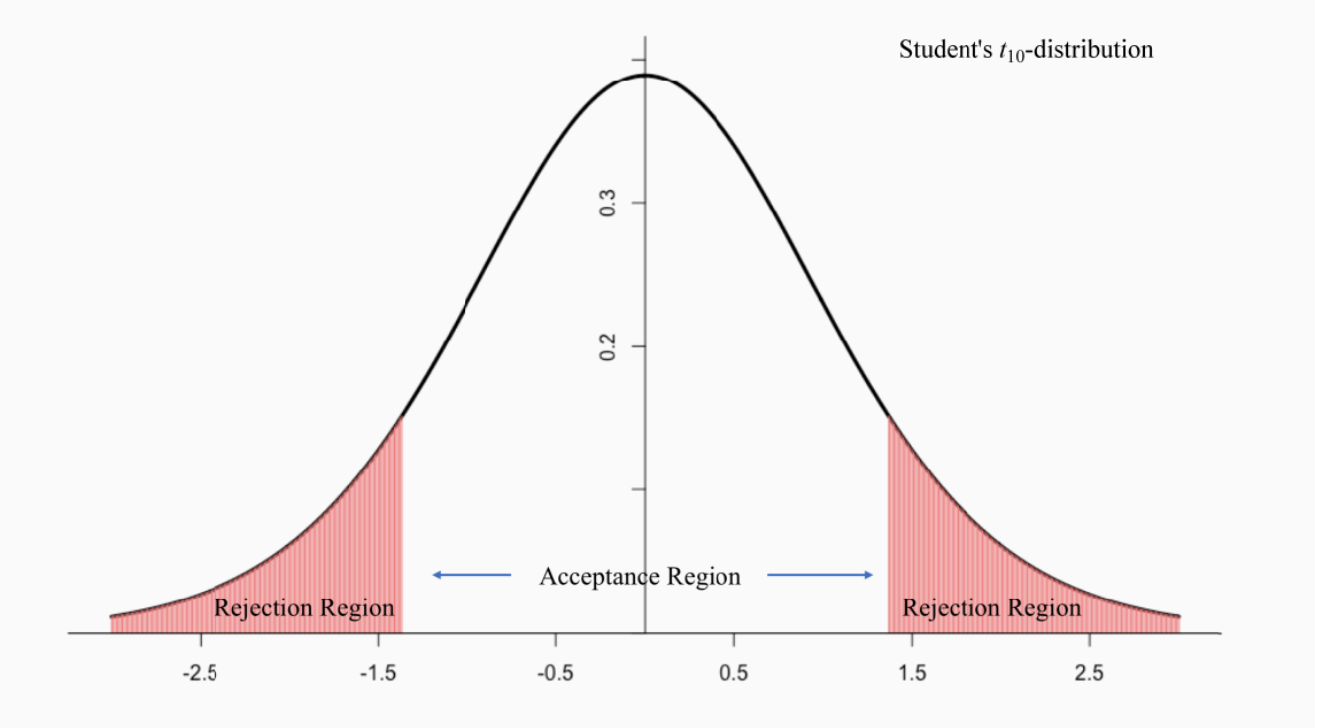


Figure 3.3 Acceptance and rejection regions of a t-test.

- Decision Rule Based on the P -value:
  - A small P-value means reject the null hypothesis.
  - A large P-value means accept the null hypothesis.

When we reject a null hypothesis, we often say there is a **statistically significant** effect. This does not mean that there is an **economic significant** effect. This is because when **large samples** are used, small and practically unimportant effects are likely to be statistically significant(since denominator of T which contains s^2 is very small when n is infinite).

Case 2: F-Test: $J > 1$, replace σ^2 by s^2

Lemma Quadratic Form of Normal Random Variables:
If a $q \times 1$ random vector $Z \sim N(0, V)$, where $V = \text{var}(Z)$ is a nonsingular $q \times q$ variance-covariance matrix, then

$$Z'V^{-1}Z \sim \chi^2_q$$

Applying this lemma, and using the result that

$$(R\hat{\boldsymbol{\beta}} - r)|\mathbf{X} \sim N[0, \sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R']$$

under $\mathbf{H}_0$, we have the quadratic form:

$$(R\hat{\boldsymbol{\beta}} - r)'[\sigma^2 R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\boldsymbol{\beta}} - r) \sim \chi^2_J$$

$$\frac{(R\hat{\beta}-r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta}-r)}{\sigma^2} \sim \chi^2_J$$

Like in constructing a t-test statistic, we should replace σ^2 by s^2 in the left hand side:

$$\frac{(R\hat{\beta}-r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta}-r)}{s^2}$$

$$\begin{aligned}
\mathrm{F} &\equiv \frac{(R\hat{\beta}-r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta}-r)/J}{s^2} \\
&= \frac{\frac{(R\hat{\beta}-r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta}-r)}{\sigma^2}/J}{\frac{(n-K)s^2}{\sigma^2}/(n-K)} \\
&\sim F_{J,n-K}
\end{aligned}$$

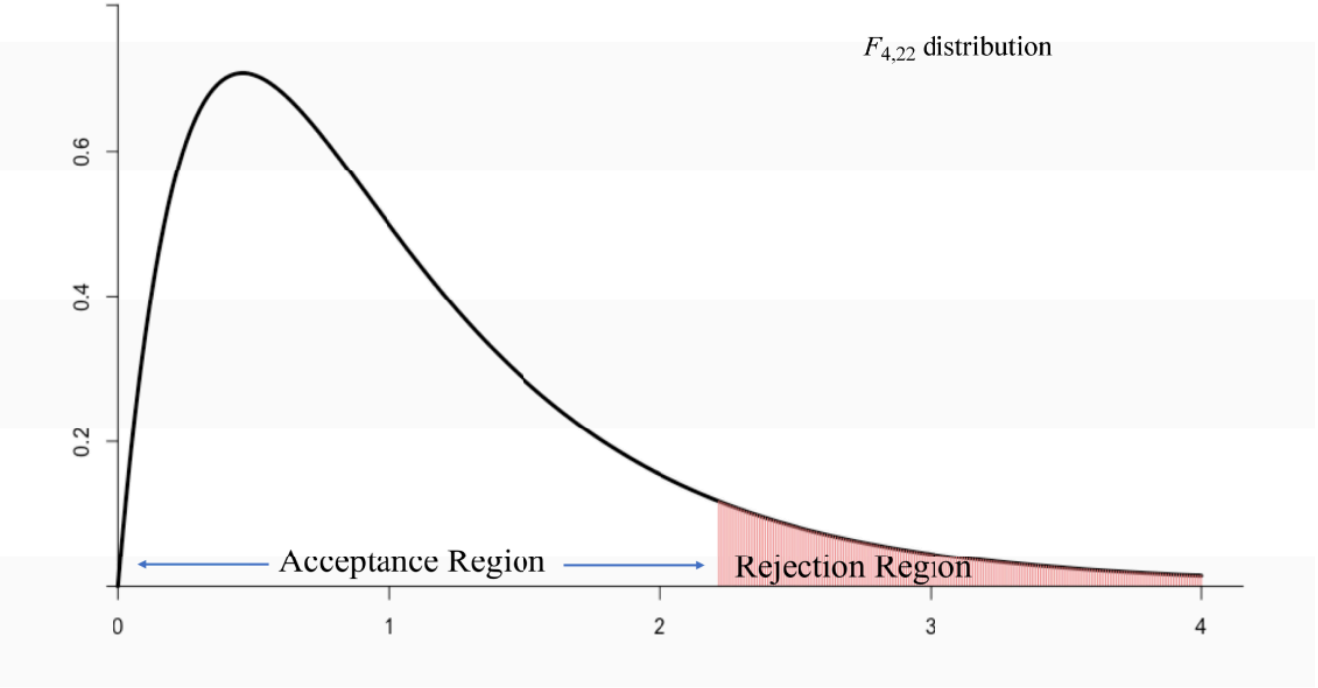


Figure 3.4 Acceptance and rejection regions of an F-test.

$$\mathrm{F} = \frac{(\tilde{e}'\tilde{e}-e'e)/J}{e'e/(n-K)} = \frac{(SSR_r - SSR_{ur})/J}{SSR_{ur}/(n-K)}$$

Wald Test:

$$W = J \cdot F = \frac{(R\hat{\beta}-r)'[R(\mathbf{X}'\mathbf{X})^{-1}R']^{-1}(R\hat{\beta}-r)}{s^2} \xrightarrow{d} \chi^2_J$$

Summary of test statistics:

$Z \Rightarrow t \to Z(when\, n \to \infty)$
$Z \Rightarrow \chi 2 \Rightarrow F \Rightarrow W \to \chi 2(when\, n \to \infty)$
$F = t^2$

Applications

- Case 1: Testing for Joint Significance of Explanatory Variables

- Case 2: Testing for Omitted Variables

  Examples: Granger Causality; Chow test
- Case 3: Testing for Linear Restrictions

GLS Estimation

Relax Ass 5 (no conditional heteroskedasticity and auto-correlation)

$$\varepsilon|\mathbf{X} \sim N(0, \sigma^2\mathbf{V})$$

It allows for conditional heteroskedasticity of known form $\sigma^2\mathbf{V} = \sigma^2 V(\mathbf{X})$. This is a strong assumption(V is often an unknown form), so GLS is not widely used in reality as OLS. Since Ass 5 is obeyed. T and F are useless.

- Unbiasedness:

$$E(\hat{\beta}|\mathbf{X}) = \beta \text{ and } E(\hat{\beta}) = \beta$$

- Variance:

$$\text{var}(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \neq \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

- Normal Distribution:

$$(\hat{\beta} - \beta)|\mathbf{X} \sim N(0, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}).$$

- Non-Zero Correlation Between $\hat{\beta}$ and e:

$$cov(\hat{\beta}, e|\mathbf{X}) \neq 0.$$

Cholesky's Decomposition:

$$\begin{aligned} \mathbf{V}^{-1} &= \mathbf{C}'\mathbf{C}, \\ \mathbf{V} &= \mathbf{C}^{-1}(\mathbf{C}')^{-1} \end{aligned}$$

Consider the original linear regression model, if we multiply the equation by C, we obtain the transformed regression model:

$$\mathbf{C}\mathbf{Y} = (\mathbf{C}\mathbf{X})\beta + \mathbf{C}\varepsilon \text{ or } Y^* = \mathbf{X}^*\beta + \varepsilon^*$$

GLS estimator:

$$\begin{aligned} \hat{\beta}^* &= (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}Y^* \\ &= (\mathbf{X}'\mathbf{C}'\mathbf{C}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{C}'\mathbf{C}Y) \\ &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}Y \end{aligned}$$

$$var(\varepsilon^*) = var(\mathbf{C}\varepsilon) = \mathbf{C}'var(\varepsilon)\mathbf{C} = \mathbf{C}'\mathbf{C}\sigma^2\mathbf{V} = \mathbf{V}^{-1}\sigma^2\mathbf{V} = \sigma^2$$

Solutions:

(1) Approach I: Adaptive Feasible GLS Estimation

(2) Approach II: White and HAC Variance-Covariance Matrix Estimation

White's (1980): heteroskedasticity consistent variance-covariance matrix estimator

Andrews (1991) and Newey and West (1987, 1994): Heteroskedasticity and Autocorrelation Consistent (HAC) variance-covariance matrix estimation

Assumptions

Ass 2.1: Linearity

$$y_i = x_i'\beta + \varepsilon_i$$

Ass 2.2: Ergodic Stationarity (for $\{y_i, x_i\}$)
Ass 2.3: Orthogonality

$$E(g_i) = E(x_i\varepsilon_i) = 0$$

Ass 2.4: Rank condition
$E(x_i x_i') = \sum_{xx}$ is full rank(non-singular).
Ass 2.5: g_i is a m.d.s with finite second moments

$$g_i \sim m.d.s$$

$E(g_i g_i')$ is nonsingular.

Consistency of OLS

Under Ass 2.1-2.4:

$$b - \beta \xrightarrow{p} 0 \quad \Rightarrow \quad b \xrightarrow{p} \beta$$

Proof:

$$b = \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} \sum_{i=1}^{n} x_i y_i$$

$$b = \left(\sum_{i=1}^{n} x_i x_i'\right)^{-1} \sum_{i=1}^{n} x_i (x_i'\beta + \varepsilon_i)$$

using Ass 2.1.

$$b = \beta + \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \cdot \frac{1}{n}\sum_{i=1}^{n} x_i \varepsilon_i$$

By Weak Law of Large Numbers(WLLN):

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i' \xrightarrow{p} E[x_i x_i'] = \Sigma_{xx}$$

using Ass 2.4, $\sum_{xx}$ is invertible.

$$\frac{1}{n}\sum_{i=1}^{n} x_i\varepsilon_i \xrightarrow{p} E[x_i\varepsilon_i] = 0$$

using Ass 2.3.

By Lemma 2.3(Continuous Mapping Theorem):

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i' \cdot \frac{1}{n}\sum_{i=1}^{n} x_i\varepsilon_i \xrightarrow{p} \Sigma_{xx}^{-1} \cdot 0 = 0$$

hence:

$$b - \beta \xrightarrow{p} 0 \quad \Rightarrow \quad b \xrightarrow{p} \beta$$

Asymptotic normality of OLS

Under Ass 2.1-2.5:

$$\sqrt{n}(b-\beta) \xrightarrow{d} N\left(0,\ \Sigma_{xx}^{-1} E(g_i g_i')\Sigma_{xx}^{-1}\right)$$

Proof:

$$\sqrt{n}(b-\beta) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \cdot \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\varepsilon_i\right)$$

let:

$$S_{xx} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i'$$

$$\bar{g} = \frac{1}{n}\sum_{i=1}^{n} x_i\varepsilon_i = \frac{1}{n}\sum_{i=1}^{n} g_i$$

so that:

$$\sqrt{n}(b-\beta) = S_{xx}^{-1} \cdot \sqrt{n}\,\bar{g}$$

By WLLN:

$$S_{xx} \xrightarrow{p} E(x_i x_i') = \Sigma_{xx}$$

By Ass2.5 + Central Limit Theorem(CLT):

$$\sqrt{n}\,\bar{g} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i \xrightarrow{d} N(0, E(g_i g_i'))$$

using Ass 2.4:

$$S_{xx}^{-1} \xrightarrow{p} \Sigma_{xx}^{-1}$$

$$\sqrt{n}\,\bar{g} \xrightarrow{d} N(0, E(g_i g_i'))$$

By Slutsky theorem(Lemma 2.4):

$$\sqrt{n}(b-\beta) = S_{xx}^{-1} \cdot \sqrt{n}\,\bar{g} \xrightarrow{d} \Sigma_{xx}^{-1} \cdot N(0, E(g_i g_i'))$$

hence:

$$\sqrt{n}(b-\beta) \xrightarrow{d} N\left(0,\ \Sigma_{xx}^{-1} E(g_i g_i') \Sigma_{xx}^{-1}\right)$$

Asymptotic variance estimator

1. $s^2 \xrightarrow{p} \sigma^2$

$$e_i = \varepsilon_i - x_i'(b-\beta)$$

$$\frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 - 2(b-\beta)' \cdot \frac{1}{n}\sum_{i=1}^{n} x_i \varepsilon_i + (b-\beta)' \cdot \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right) \cdot (b-\beta)$$

- For the first term in RHS:

$$\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 \xrightarrow{p} E(\varepsilon_i^2) = \sigma^2$$

using Ass 2.2 + WLLN.

- For the second term:

$$\frac{1}{n}\sum_{i=1}^{n} x_i \varepsilon_i = \bar{g} \xrightarrow{p} E(g_i) = 0$$

using Ass 2.3 + WLLN.

$b - \beta \xrightarrow{p} 0$ (consistency)
so $(b-\beta)' \cdot \frac{1}{n}\sum_{i=1}^{n} x_i \varepsilon_i \xrightarrow{p} 0$

- For the third term:

$$\frac{1}{n}\sum_{i=1}^{n} x_i x_i' \xrightarrow{p} E(x_i x_i') \equiv \Sigma_{xx}$$

and $b - \beta \xrightarrow{p} 0$
so $(b-\beta)' \cdot \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right) \cdot (b-\beta) \xrightarrow{p} 0$

such that:

$$p\lim \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \sigma^2$$

$$s^2 = \frac{1}{n-K} \cdot \sum_{i=1}^{n} e_i^2 = \frac{1}{n} \sum_{i=1}^{n} e_i^2$$

hence:

$$s^2 \xrightarrow{p} \sigma^2$$

2. $\hat{S} \xrightarrow{p} S$

$$S = E(g_i g_i') = E(x_i \varepsilon_i \varepsilon_i' x_i')$$

By the law of iterated expectation(IIE) and under conditional homoskedasticity:

$$S = E(E(x_i \varepsilon_i \varepsilon_i' x_i' | x_i) = E(x_i x_i' E(\varepsilon_i \varepsilon_i' | x_i)) = \sigma^2 E(x_i x_i')$$

for: $s^2 \xrightarrow{p} \sigma^2$, $S_{xx} \xrightarrow{p} E(x_i x_i') = \Sigma_{xx}$
then: $\hat{S} \xrightarrow{p} S$

3. $\widehat{\text{Avar}}(b) \xrightarrow{p} \text{Avar}(b)$

$$\widehat{\text{Avar}}(b) = S_{xx}^{-1} \hat{S} S_{xx}^{-1} \xrightarrow{p} \Sigma_{xx}^{-1} E(g_i g_i') \Sigma_{xx}^{-1} = \text{Avar}(b)$$

Hypothesis testing

Testing linear hypotheses

the same as chap1, see case 1 below.

Testing non-linear hypotheses

$H_0 : h(b_{OLS}) = 0$

Use the **delta method**: $H(b) \equiv \frac{\partial h(b)}{\partial b'}$

$$W = n \cdot h(b_{OLS})' \left[H(b_{OLS}) Avar(b_{OLS}) H(b_{OLS})' \right]^{-1} h(b_{OLS}) \xrightarrow{d} \chi^2(p)$$

p is the number of constraint conditions.
Detail of Wald test will be discussed in chap3.

Summary of Yongmiao Hong's book(chap4):

$$(1)\widehat{\beta} \xrightarrow{p} \beta \, as \, n \to \infty$$
$$(2)\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} N(0, Q^{-1}VQ^{-1}) \, as \, n \to \infty$$
$$where \, Q = E(X_i X_i'), V = E(X_i X_i' \varepsilon_i^2)$$
$$(3)if \, E(\varepsilon_i^2 | X_i) = \sigma^2, \, V = \sigma^2 Q, \, Q^{-1}VQ^{-1} = \sigma^2 Q^{-1}$$
$$Q : \widehat{Q} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i' \xrightarrow{p} Q$$
$$V : \widehat{V} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i' e_i^2 \xrightarrow{p} V$$
$$\widehat{Q}^{-1}\widehat{V}\widehat{Q}^{-1} \xrightarrow{p} Q^{-1}VQ^{-1}$$

Case I: Conditional Homoskedasticity

$J = 1$:

$$T = \frac{R\hat{\beta} - r}{\sqrt{s^2 R(\mathbf{X}'\mathbf{X})^{-1}R'}} \xrightarrow{d} N(0,1)$$

$J > 1$:

$$W \equiv (R\hat{\beta} - r)' \left[s^2 R(\mathbf{X}'\mathbf{X})^{-1}R'\right]^{-1}(R\hat{\beta} - r) = J \cdot F \xrightarrow{d} \chi_J^2$$

Case II: Conditional Heteroskedasticity

$J = 1$:

$$T_r = \frac{\sqrt{n}(R\hat{\beta} - r)}{\sqrt{R\hat{Q}^{-1}\hat{V}\hat{Q}^{-1}R'}} \xrightarrow{d} N(0,1)$$

$J > 1$:

$$W_r = n(R\hat{\beta} - r)' \left(R\hat{Q}^{-1}\hat{V}\hat{Q}^{-1}R'\right)^{-1}(R\hat{\beta} - r) \xrightarrow{d} \chi_J^2$$

MLE Basics

$$y_i = x_i'\beta + u_i, \quad u_i \sim NID(0, \sigma^2)$$

normal distribution:

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

the loglikelihood function:

$$\log L(\beta,\sigma^2) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{N}\frac{(y_i - x_i'\beta)^2}{\sigma^2}$$

the score vector:

$$s(\theta) \equiv \frac{\partial \log L(\theta)}{\partial \theta} = \sum_{i=1}^{N}\frac{\partial \log L_i(\theta)}{\partial \theta} \equiv \sum_{i=1}^{N} s_i(\theta)$$

$$s_i(\beta,\sigma^2) = \begin{pmatrix} \frac{\partial \log L_i(\beta,\sigma^2)}{\partial \beta} \\ \frac{\partial \log L_i(\beta,\sigma^2)}{\partial \sigma^2} \end{pmatrix} = \begin{pmatrix} \frac{(y_i - x_i'\beta)}{\sigma^2}x_i \\ -\frac{1}{2\sigma^2} + \frac{1}{2}\frac{(y_i - x_i'\beta)^2}{\sigma^4} \end{pmatrix}$$

FOC implies:

$$\hat{\beta} = \left(\sum_{i=1}^{N} x_i x_i'\right)^{-1}\sum_{i=1}^{N} x_i y_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - x_i'\hat{\beta})^2$$

Information matrix:

$$I(\theta) \equiv E[s(\theta)s(\theta)'] = -E[\frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}]$$

$$I_i(\beta,\sigma^2) = E[s_i(\beta,\sigma^2)s_i(\beta,\sigma^2)']$$

calculation:

$$E(s_i s_i') = E[\begin{matrix} \frac{u_i}{\sigma^2}\cdot x_i \\ -\frac{1}{2\sigma^2} + \frac{u_i^2}{2\sigma^4} \end{matrix}] \cdot [\frac{u_i}{\sigma^2}x_i, -\frac{1}{2\sigma^2} + \frac{u_i^2}{2\sigma^4}]$$

$$= E\begin{bmatrix} \left(\frac{u_i}{\sigma^2}x_i\right)^2 & -\frac{u_i x_i}{2\sigma^4} + \frac{u_i^3 x_i}{2\sigma^6} \\ -\frac{u_i x_i}{2\sigma^4} + \frac{u_i^3 x_i}{2\sigma^6} & \left(-\frac{1}{2\sigma^2} + \frac{u_i^2}{2\sigma^4}\right)^2 \end{bmatrix}$$

since $E(u_i) = 0$, $E(u_i^2) = \sigma^2$, $E(u_i^3) = 0$, $E(u_i^4) = 3\sigma^4$, we can finally get

$$I_i(\beta,\sigma^2)=\begin{pmatrix}\frac{1}{\sigma^2}x_ix_i' & 0\\ 0 & \frac{1}{2\sigma^4}\end{pmatrix}$$

The asymptotic covariance matrix:

$$V=I(\beta,\sigma^2)^{-1}=\begin{pmatrix}\sigma^2\Sigma_{xx}^{-1} & 0\\ 0 & 2\sigma^4\end{pmatrix}$$

$$\sqrt{N}(\hat{\beta}-\beta)\to\mathcal{N}(0,\sigma^2\Sigma_{xx}^{-1})$$

$$\sqrt{N}(\hat{\sigma}^2-\sigma^2)\to\mathcal{N}(0,2\sigma^4)$$

Hypothesis test

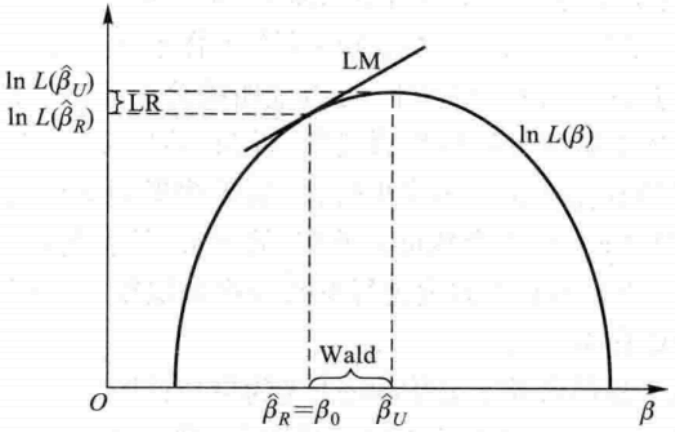


图 6.6　三类渐近等价的统计检验

LR(Likelihood Ratio Test)

$$\log L=-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log(\tilde{\sigma}^2)-\frac{1}{2\tilde{\sigma}^2}\tilde{u}'\tilde{u}$$

$$SSR=\tilde{u}'\tilde{u},\ \tilde{\sigma}^2=\frac{SSR}{n}$$

$$\Rightarrow logL=-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log\frac{SSR}{n}-\frac{n}{2}$$

$$\Rightarrow L=(\frac{2\pi}{n})^{-\frac{n}{2}}\cdot e^{-\frac{n}{2}}\cdot(SSR)^{-\frac{n}{2}}$$

$$\lambda\equiv L_U/L_R\equiv(\frac{SSR_U}{SSR_R})^{-n/2}$$

$$F=\frac{(SSR_R-SSR_U)/\#r}{SSR_U/n-k-1}=\frac{n-k-1}{\#r}(\lambda^{\frac{2}{n}}-1)$$

Wald Test

We have already seen in chap 2:

$$W \equiv (R\tilde{\beta} - r)' \left[s^2 R(\mathbf{X}'\mathbf{X})^{-1} R' \right]^{-1} (R\tilde{\beta} - r) \xrightarrow{d} \chi^2_J$$

$$W = n \cdot h(b_{OLS})' \left[H(b_{OLS}) Avar(b_{OLS}) H(b_{OLS})' \right]^{-1} h(b_{OLS}) \xrightarrow{d} \chi^2(p)$$

For nonlinear hypotheses:

$$H_0 : h(\theta) = 0$$

By first-order Taylor expansion:

$$h(\hat{\theta}) \approx H(\theta)(\hat{\theta} - \theta)$$

$$H(\theta) = \frac{\partial h(\theta)}{\partial \theta'}$$

Wald Test Statistic:

$$W = n \cdot h(\hat{\theta})' [\hat{H} \widehat{Avar}(\hat{\theta}) \hat{H}']^{-1} h(\hat{\theta})$$

$\widehat{Avar}(\hat{\theta})$ can be the inverse of information matrix.

Wald Test Examples(see lecture note):
(1)Tests of Exclusion Restrictions
(2)Tests of Statistical Significance
(3)Tests of Nonlinear Restriction

Examples

Verbeek 6.1.1+6.2.3:

(1)

$$L(p) = \binom{N}{N_1} p^{N_1} (1-p)^{N-N_1}$$

$$\log L(p) = \log \binom{N}{N_1} + N_1 \log p + (N - N_1) \log(1-p)$$

(2)

$$\frac{d \log L(p)}{dp} = \frac{N_1}{p} - \frac{N - N_1}{1-p} = 0$$

$$\hat{p} = \frac{N_1}{N}$$

(3)

$$\log L_i(p) = y_i \log p + (1 - y_i) \log(1-p)$$

$$\frac{\partial \log L_i(p)}{\partial p} = \frac{y_i}{p} - \frac{1 - y_i}{1-p}$$

$$-\frac{\partial^2 \log L_i(p)}{\partial p^2} = \frac{y_i}{p^2} + \frac{1-y_i}{(1-p)^2}$$

$$I = E\left\{-\frac{\partial^2 \log L_i(p)}{\partial p^2}\right\} = \frac{E\{y_i\}}{p^2} + \frac{1-E\{y_i\}}{(1-p)^2} = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}$$

note that $E\{y_i\} = p$

$$V = I^{-1}$$

$$\sqrt{n}(\hat{p} - p) \to \mathcal{N}(0, p(1-p))$$

(4)

$$p \pm z_{a/2}\sqrt{\frac{p(1-p)}{n}} = 0.44 \pm 1.96\sqrt{\frac{0.44(1-0.44)}{100}} = (0.3427, 0.5373)$$

(5)$H_0 : \hat{p} = p_0$ for a given value p_0
Wald test :

$$W = N(\hat{p} - p_0)[\hat{p}(1-\hat{p})]^{-1}(\hat{p} - p_0)$$

For LR test we need to compare the unrestricted and the restricted model:

$$\log L(\hat{p}) = N_1 \log(N_1/N) + (N - N_1)\log(1 - N_1/N)$$

$$\log L(\tilde{p}) = N_1 \log(p_0) + (N - N_1)\log(1 - p_0)$$

$$LR = 2(\log L(\hat{p}) - \log L(\tilde{p}))$$

Verbeek Exercise 6.2:

a.

$$L(\beta_1, \beta_2) = \prod_{i=1}^{N} \frac{e^{-e^{\beta_1+\beta_2 x_i}} e^{\beta_1+\beta_2 x_i y_i}}{y_i!}$$

$$\log L(\beta_1, \beta_2) = \sum_{i=1}^{N} \left[-e^{\beta_1+\beta_2 x_i} + y_i(\beta_1 + \beta_2 x_i) - \log(y_i!)\right]$$

b.

$$\frac{\partial \log L}{\partial \beta_1} = \sum_{i=1}^{N} \left(y_i - e^{\beta_1+\beta_2 x_i}\right) = \sum_{i=1}^{N} (y_i - \lambda_i)$$

$$\frac{\partial \log L}{\partial \beta_2} = \sum_{i=1}^{N} \left(y_i - e^{\beta_1+\beta_2 x_i}\right) x_i = \sum_{i=1}^{N} (y_i - \lambda_i) x_i$$

$$s(\beta) = \begin{pmatrix} \sum_{i=1}^{N} (y_i - \lambda_i) \\ \sum_{i=1}^{N} (y_i - \lambda_i) x_i \end{pmatrix}$$

$$s_i(\beta) = \begin{pmatrix} y_i - \lambda_i \\ (y_i - \lambda_i) x_i \end{pmatrix}$$

given $E[y_i|x_i] = \lambda_i$:

$$E[y_i - \lambda_i|x_i] = 0$$

$$E[(y_i - \lambda_i)x_i|x_i] = x_i E[y_i - \lambda_i|x_i] = 0$$

$$E[\boldsymbol{s}_i(\beta)|x_i] = E[\boldsymbol{s}_i(\beta)] = 0$$

c. $\frac{\partial^2 \log L}{\partial \beta_1^2} = \sum_{i=1}^N (-\lambda_i)$, $\frac{\partial^2 \log L}{\partial \beta_2^2} = \sum_{i=1}^N (-\lambda_i x_i^2)$, $\frac{\partial^2 \log L}{\partial \beta_1 \partial \beta_2} = \sum_{i=1}^N (-\lambda_i x_i)$

$$H(\beta) = -\begin{pmatrix} \sum \lambda_i & \sum \lambda_i x_i \\ \sum \lambda_i x_i & \sum \lambda_i x_i^2 \end{pmatrix}$$

$$I(\beta) = -E[H(\beta)] = \begin{pmatrix} \sum E[\lambda_i] & \sum E[\lambda_i x_i] \\ \sum E[\lambda_i x_i] & \sum E[\lambda_i x_i^2] \end{pmatrix} = \begin{pmatrix} \sum \lambda_i & \sum \lambda_i x_i \\ \sum \lambda_i x_i & \sum \lambda_i x_i^2 \end{pmatrix}$$

the asymptotic covariance matrix: $\mathrm{Var}(\beta) = I(\beta)^{-1}$
consistent estimator: $\widehat{\mathrm{Var}}(\hat{\beta}) = \hat{I}(\hat{\beta})^{-1}$, $\hat{\lambda}_i = \exp(\hat{\beta}_1 + \hat{\beta}_2 x_i)$

Assumptions

Ass 3.1: Linearity

$$y_i = z_i'\delta + \varepsilon_i$$

Ass 3.2: Ergodic Stationarity (for $w_i = \{y_i, x_i, z_i\}$)

Ass 3.3: Orthogonality

$$E(g_i) = E(x_i\varepsilon_i) = 0$$

Ass 3.4: Rank condition
$E(x_i z_i') \sum_{xz}$ is full rank, $Rank(\sum_{xz}) = L$
when K=L, $\sum_{xz}^{-1}$ exisits.

Ass 3.5: g_i is a m.d.s with finite second moments

$$g_i \sim m.d.s$$

$E(g_i g_i')$ is nonsingular.

Ass 3.6: finite fourth moments
$E[(x_{ik}z_{il}')^2]$ exists and finite.

Ass 3.7: Conditional Homoskedasticity

2SLS

$$Z = X\pi + v$$

first stage:

$$\hat{Z} = X(X'X)^{-1}X'Z = PZ$$

second stage:

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

By WLLN, CLT and some lemmas(see GMM):

$$\hat{\delta}_{IV} \xrightarrow{p} \delta$$

CAN: Consistent and Asymptotically Normal

GMM Basics

$$y_i = z_i'\delta + \varepsilon_i$$

$$E(g_i) \equiv E(x_i\varepsilon_i) = E(x_i(y_i - z_i'\delta)) = 0$$

$$g_n(\tilde{\delta}) = \frac{1}{n}\sum_{i=1}^{n} x_i(y_i - z_i'\tilde{\delta}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i z_i'\tilde{\delta} = s_{xy} - S_{xz}\tilde{\delta}$$

GMM estimator:

$$\tilde{\delta}(\hat{W}) = \arg\min_{\tilde{\delta}}\ n\, g_n(\tilde{\delta})'\hat{W} g_n(\tilde{\delta}) = \arg\min_{\tilde{\delta}}\ n(s_{xy} - S_{xz}\tilde{\delta})'\hat{W}(s_{xy} - S_{xz}\tilde{\delta})$$

FOC:

$$\tilde{\delta}(\hat{W}) = \left(S_{xz}'\hat{W}S_{xz}\right)^{-1} S_{xz}'\hat{W}s_{xy}$$

sample error:

$$\tilde{\delta}(\hat{W}) - \delta = \left(S_{xz}'\hat{W}S_{xz}\right)^{-1} S_{xz}'\hat{W}s_{xy} - \left(S_{xz}'\hat{W}S_{xz}\right)^{-1}\left(S_{xz}'\hat{W}S_{xz}\right)\delta = (S_{xz}'\hat{W}S_{xz})^{-1} S_{xz}'\hat{W}\bar{g}$$

base on

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i(z_i'\delta + \varepsilon_i) = \frac{1}{n}\sum_{i=1}^{n} x_i z_i'\delta + \frac{1}{n}\sum_{i=1}^{n} x_i\varepsilon_i = S_{xz}\delta + \bar{g}$$

Assumptions

the same as chap4

Proposition 3.1 Asymptotic distribution of GMM estimator

Consistency

Under Ass 3.1-3.4:

$$p\lim_{n\to\infty} \tilde{\delta}(\hat{W}) = \delta$$

Proof:

Ass 3.2, 3.4 ⇒$S_{xz} \xrightarrow{p} \Sigma_{xz}$, Σ_{xz} is column full rank$(K \geq L)$

Ass 3.2-3.3 ⇒$\bar{g} = \frac{1}{n}\sum_{i=1}^{n} g_i \xrightarrow{p} 0$

$\hat{W} \xrightarrow{p} W$ by the definition of W(**symmetric** and **positive definite** weight matrix)

By Slutsky theorem:

$$S'_{xz}\hat{W}S_{xz} \xrightarrow{p} \Sigma'_{xz}W\Sigma_{xz}$$

$$S'_{xz}\hat{W}\bar{g} \xrightarrow{p} \Sigma'_{xz}W \cdot 0 = 0$$

hence:

$$\tilde{\delta}(\hat{W}) - \delta \xrightarrow{p} \left(\Sigma'_{xz}W\Sigma_{xz}\right)^{-1} \cdot 0 = 0$$

$$p\lim \tilde{\delta}(\hat{W}) = \delta$$

Asymptotic normality

Under Ass 3.5:

$$\sqrt{n}\left(\tilde{\delta}(\hat{W}) - \delta\right) \xrightarrow{d} N\left(0,\ \mathrm{Avar}(\tilde{\delta}(\hat{W}))\right)$$

$$\mathrm{Avar}(\tilde{\delta}(\hat{W})) = \left(\Sigma'_{xz}W\Sigma_{xz}\right)^{-1}\Sigma'_{xz}WSW\Sigma_{xz}\left(\Sigma'_{xz}W\Sigma_{xz}\right)^{-1}$$

$$S = E[g_i g_i'] = E[x_i x_i' \varepsilon_i^2]$$

proof:

$$\sqrt{n}\left(\tilde{\delta}(\hat{W}) - \delta\right) = \left(S'_{xz}\hat{W}S_{xz}\right)^{-1}S'_{xz}\hat{W} \cdot \sqrt{n}\,\bar{g}$$

By CLT:

$$\sqrt{n}\,\bar{g} \xrightarrow{d} N(0, S)$$

By Slutsky theorem:

$$S'_{xz}\hat{W}S_{xz} \xrightarrow{p} \Sigma'_{xz}W\Sigma_{xz}$$

$$S'_{xz}\hat{W} \xrightarrow{p} \Sigma'_{xz}W$$

hence:

$$\sqrt{n}\left(\tilde{\delta}(\hat{W}) - \delta\right) \xrightarrow{d} \left(\Sigma'_{xz}W\Sigma_{xz}\right)^{-1}\Sigma'_{xz}W \cdot N(0, S)$$

$$\mathrm{Avar}(\tilde{\delta}(\hat{W})) = \left(\Sigma'_{xz}W\Sigma_{xz}\right)^{-1}\Sigma'_{xz}WSW\Sigma_{xz}\left(\Sigma'_{xz}W\Sigma_{xz}\right)^{-1}$$

Consistent estimate of $\mathrm{Avar}(\tilde{\delta}(\hat{W}))$

$$\widehat{\mathrm{Avar}}(\tilde{\delta}(\hat{W})) \xrightarrow{p} \mathrm{Avar}(\tilde{\delta}(\hat{W}))$$

Proof:

$$\widehat{\text{Avar}}(\tilde{\delta}(\hat{W})) = (S_{xz}'\hat{W}S_{xz})^{-1}S_{xz}'\hat{W}\hat{S}\hat{W}S_{xz}(S_{xz}'\hat{W}S_{xz})^{-1}$$

$$S_{xz} \xrightarrow{p} \Sigma_{xz}, \quad \hat{W} \xrightarrow{p} W, \quad \hat{S} \xrightarrow{p} S$$

$$\widehat{\text{Avar}}(\tilde{\delta}(\hat{W})) \xrightarrow{p} (\Sigma_{xz}'W\Sigma_{xz})^{-1}\Sigma_{xz}'WSW\Sigma_{xz}(\Sigma_{xz}'W\Sigma_{xz})^{-1}$$

Proposition 3.2 consistent estimation of error variance

Under Ass 3.1-3.2 + $E(z_i z_i')$ exists and is finite:

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 \xrightarrow{p} E(\varepsilon_i^2)$$

Proof:

$$\hat{\varepsilon}_i = \varepsilon_i - z_i'(\hat{\delta} - \delta)$$

$$\hat{\varepsilon}_i^2 = \varepsilon_i^2 - 2(\hat{\delta} - \delta)'z_i\varepsilon_i + (\hat{\delta} - \delta)'z_i z_i'(\hat{\delta} - \delta)$$

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 - 2(\hat{\delta} - \delta)'\left(\frac{1}{n}\sum_{i=1}^{n}z_i\varepsilon_i\right) + (\hat{\delta} - \delta)'\left(\frac{1}{n}\sum_{i=1}^{n}z_i z_i'\right)(\hat{\delta} - \delta)$$

- For the first term in RHS, using Ass 3.2 + WLLN: $\frac{1}{n}\sum \varepsilon_i^2 \xrightarrow{p} E(\varepsilon_i^2)$
- For the second term, $\hat{\delta} - \delta \to 0$ (consistency), $\frac{1}{n}\sum z_i\varepsilon_i \xrightarrow{p} E(z_i\varepsilon_i)$ exists and is finite (by Cauchy-Schwartz inequality)
- For the third term, $\hat{\delta} - \delta \to 0$, $\frac{1}{n}\sum z_i z_i' \xrightarrow{p} E(z_i z_i')$ exists and is finite

hence:

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 \xrightarrow{p} E(\varepsilon_i^2)$$

The Efficient GMM Estimator

To minimize $\text{Avar}(\tilde{\delta}(\hat{W}))$, choose $\hat{S}^{-1}$ as the weighting matrix.

$$Avar(\tilde{\delta}(\hat{W})) = (\Sigma_{xz}'W\Sigma_{xz})^{-1}\Sigma_{xz}'WSW\Sigma_{xz}(\Sigma_{xz}'W\Sigma_{xz})^{-1}$$

$$Avar(\tilde{\delta}(\hat{S}^{-1})) = (\Sigma_{xz}'S^{-1}\Sigma_{xz})^{-1}$$

$$\widehat{Avar}(\tilde{\delta}(\hat{S}^{-1})) = (S_{xz}'\hat{S}^{-1}S_{xz})^{-1}$$

From GMM to 2SLS

Under Ass 3.7(conditional homoskedasticity):

$$S = E(x_i x_i' \varepsilon_i^2) = E[x_i x_i' E(\varepsilon_i^2 | x_i)] = \sigma^2 E(x_i x_i')$$

as is proofed in chap2:
$s^2 \xrightarrow{p} \sigma^2$, $S_{xx} \xrightarrow{p} E(x_i x_i') = \Sigma_{xx}$, $\hat{S} \xrightarrow{p} S$

$$\begin{aligned}
\hat{\beta}_{GMM} &= (S_{xz}' \hat{S}^{-1} S_{xz})^{-1} S_{xz}' \hat{S}^{-1} s_{xy} \\
&= \left(S_{xz}' (s^2 S_{xx})^{-1} S_{xz}\right)^{-1} S_{xz}' (s^2 S_{xx})^{-1} s_{xy} \\
&= \left(S_{xz}' S_{xx}^{-1} S_{xz}\right)^{-1} S_{xz}' S_{xx}^{-1} s_{xy} \\
&= \left(\frac{Z'X}{n} (\frac{X'X}{n})^{-1} \frac{X'Z}{n} \right)^{-1} \frac{Z'X}{n} (\frac{X'X}{n})^{-1} \frac{X'y}{n} \\
&= \left(Z'X(X'X)^{-1}X'Z\right)^{-1} Z'X(X'X)^{-1}X'y \\
&= \left(Z'P_X P_X Z\right)^{-1} Z'P_X y \\
&= (\hat{Z}'\hat{Z})^{-1}\hat{Z}'y \\
&= \hat{\beta}_{2SLS}
\end{aligned}$$

don't forget the projection matrix P is symmetric and idempotent.

Proposition 3.3 Robust t ratio and Wald statistics

Under Ass 3.1-3.5:

(a) Under the null $H_0 : \delta = \bar{\delta}$

$$t \equiv \frac{\sqrt{n}(\tilde{\delta}(\hat{W}) - \bar{\delta})}{\sqrt{\widehat{Avar}(\tilde{\delta}(\hat{W}))}} \xrightarrow{d} N(0,1)$$

(b) Under the null $H_0 : R\delta = r$

$$W \equiv n(R\tilde{\delta}(\hat{W}) - r)' \{R[\widehat{Avar}(\tilde{\delta}(\hat{W}))]R'\}^{-1}(R\tilde{\delta}(\hat{W}) - r) \xrightarrow{d} \chi^2(\#r)$$

(c) Under the null $H_0 : a(\delta) = 0$

$$W \equiv n\, a(\tilde{\delta}(\hat{W}))' \{A(\tilde{\delta}(\hat{W}))[\widehat{Avar}(\tilde{\delta}(\hat{W}))]A(\tilde{\delta}(\hat{W}))'\}^{-1} a(\tilde{\delta}(\hat{W})) \xrightarrow{d} \chi^2(\#a)$$

(d) $\bar{\delta}$ is from restricted model, $\tilde{\delta}$ is from unrestricted model:

$$LR \equiv J(\bar{\delta}(\hat{S}^{-1}), \hat{S}^{-1}) - J(\tilde{\delta}(\hat{S}^{-1}), \hat{S}^{-1}) \xrightarrow{d} \chi^2$$

Proposition 3.6 Hansen's test of overidentifying restrictions

Under Ass 3.1-3.5:

$$J(\tilde{\delta}(\hat{S}^{-1}), \hat{S}^{-1}) = n g_n(\tilde{\delta}(\hat{S}^{-1}))' \hat{S}^{-1} g_n(\tilde{\delta}(\hat{S}^{-1})) \xrightarrow{d} \chi^2(K-L)$$

$\sqrt{n}\bar{g} \xrightarrow{d} N(0, \mathrm{S}), \hat{S} \xrightarrow{p} S$

Under conditional homoskedasticity, J becomes **Sargan Statistic**:

$$n(s_{xy} - S_{xz}\tilde{\delta}_{2SLS})'(s^2 S_{xx})^{-1}(s_{xy} - S_{xz}\tilde{\delta}_{2SLS})$$

Proposition 3.7 Testing a subset of orthogonality conditions

Under Ass 3.1-3.5:

$$C \equiv J - J_1 \to \chi^2(K - K_1)$$

need $K_1 \geq L$

Basic formulars:

$$g_n(\delta) = \bar{g} = \frac{1}{n}\sum_{i=1}^{n} g_i$$

GMM estimator:

$$\tilde{\delta}(\hat{W}) = \left(S_{xz}'\hat{W}S_{xz}\right)^{-1} S_{xz}'\hat{W}s_{xy}$$

sample error:

$$\tilde{\delta} - \delta = (S_{xz}'\hat{W}S_{xz})^{-1} S_{xz}'\hat{W}\bar{g}$$

sample moment:

$$g_n(\tilde{\delta}) = s_{xy} - S_{xz}\tilde{\delta}$$

In **efficient GMM**: $\hat{W} = \hat{S}^{-1}$. By the definition, $\hat{W}$ is a $K \times K$ symmetric and positive definite matrix.

First-order Taylor expansion:

$$\begin{aligned} g_n(\tilde{\delta}) &\approx g_n(\delta) + \frac{\partial g_n(\tilde{\delta})}{\partial \tilde{\delta}}\Big|_{\tilde{\delta}=\delta} (\tilde{\delta} - \delta) \\ &= \bar{g} - S_{xz}(\tilde{\delta} - \delta) \\ &= \bar{g} - S_{xz}(S_{xz}'\hat{W}S_{xz})^{-1} S_{xz}'\hat{W}\bar{g} \end{aligned}$$

the weighted form projection matrix:

$$P = S_{xz}(S_{xz}'\hat{W}S_{xz})^{-1} S_{xz}'\hat{W}$$

$Rank(P) = L$, since S_{xz} is full column rank(=L).

$$g_n(\tilde{\delta}) = (I_K - P)\bar{g} = M g_n(\delta)$$

$$Rank(M) = tr(M) = tr(I_K) - tr(P) = K - L$$

Conclusion:

$$J(\delta, S^{-1}) = n.\,\bar{g}' S^{-1}\bar{g} = (\sqrt{n}\bar{g})' S^{-1}(\sqrt{n}\bar{g}) \sim \chi^2(K)$$

$$J(\tilde{\delta}(\hat{S}^{-1}), \hat{S}^{-1}) = n.\, g_n(\tilde{\delta}(\hat{S}^{-1}))' \hat{S}^{-1} g_n(\tilde{\delta}(\hat{S}^{-1})) \xrightarrow{d} \chi^2(K-L)$$

The reason for the loss of degrees of freedom is: When constructing $g_n(\tilde{\delta})$, we used $\tilde{\delta}$ to replace the true δ, which is equivalent to imposing L constraints on the sample moments (by projecting $\bar{g}$ onto the column space of S_{xz} through the projection matrix P), which can also explain why we use the degrees of freedom $n - K$ in OLS estimation.

Assumptions

Ass 4.1: Linearity

$$y_{im} = \mathbf{z}_{im}^{'}\delta_m + \varepsilon_{im}$$

$$\begin{bmatrix} \mathbf{y}_1 \\ n \times 1 \\ \vdots \\ \mathbf{y}_M \\ n \times 1 \end{bmatrix} = \begin{bmatrix} \mathbf{Z}_1 & & \\ n \times L_1 & & \\ & \ddots & \\ & & \mathbf{Z}_M \\ & & n \times L_M \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta}_1 \\ L_1 \times 1 \\ \vdots \\ \boldsymbol{\delta}_M \\ L_M \times 1 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ n \times 1 \\ \vdots \\ \boldsymbol{\varepsilon}_M \\ n \times 1 \end{bmatrix}$$

Ass 4.2: Jointly Ergodic Stationarity

$$\{\mathbf{w}_i\} = \{y_{i1}, \ldots, y_{iM}, \mathbf{z}_{i1}, \ldots, \mathbf{z}_{iM}, \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iM}\}$$

Ass 4.3: Orthogonality

$$E(\mathbf{x}_{im} \cdot \varepsilon_{im}) = \mathbf{0}$$

We don't assume cross equation orthogonalities.

Ass 4.4: Rank condition
$E(\mathbf{x}_{im}\mathbf{z}_{im}')$ is full column rank.

Ass 4.5: g_i is a m.d.s with finite second moments

$$g_i \sim m.d.s$$

$E(g_i g_i')$ is nonsingular.

Ass 4.6: finite fourth moments
$E[(x_{imk}z_{ihl}')^2]$ exists and finite.

Ass 4.7: conditional homoskedasticity

MEGMM

$$\bar{\mathbf{g}} \equiv \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{i1}\cdot\varepsilon_{i1} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{iM}\cdot\varepsilon_{iM} \end{bmatrix} = \mathbf{g}_n(\boldsymbol{\delta})$$

population moment:

$$\begin{aligned} E[\mathbf{g_i}] &\equiv \begin{bmatrix} E[\mathbf{x}_{i1}\cdot(y_{i1}-\mathbf{z}_{i1}^{'}\tilde{\boldsymbol{\delta}}_1)] \\ \vdots \\ E[\mathbf{x}_{iM}\cdot(y_{iM}-\mathbf{z}_{iM}^{'}\tilde{\boldsymbol{\delta}}_M)] \end{bmatrix} \\ &= \begin{bmatrix} E(\mathbf{x}_{i1}\cdot y_{i1}) \\ \vdots \\ E(\mathbf{x}_{iM}\cdot y_{iM}) \end{bmatrix} - \begin{bmatrix} E(\mathbf{x}_{i1}\mathbf{z}_{i1}^{'})\tilde{\boldsymbol{\delta}}_1 \\ \vdots \\ E(\mathbf{x}_{iM}\mathbf{z}_{iM}^{'})\tilde{\boldsymbol{\delta}}_M \end{bmatrix} \\ &= \begin{bmatrix} E(\mathbf{x}_{i1}\cdot y_{i1}) \\ \vdots \\ E(\mathbf{x}_{iM}\cdot y_{iM}) \end{bmatrix} - \begin{bmatrix} E(\mathbf{x}_{i1}\mathbf{z}_{i1}^{'}) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & E(\mathbf{x}_{iM}\mathbf{z}_{iM}^{'}) \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{\delta}}_1 \\ \vdots \\ \tilde{\boldsymbol{\delta}}_M \end{bmatrix} \\ &\equiv \boldsymbol{\sigma}_{xy} - \boldsymbol{\Sigma}_{xz}\tilde{\boldsymbol{\delta}} \end{aligned}$$

sample moment:

$$\begin{aligned} \mathbf{g}_n(\tilde{\boldsymbol{\delta}}) &= \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{i1}\cdot(y_{i1}-\mathbf{z}_{i1}^{'}\tilde{\boldsymbol{\delta}}_1) \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{iM}\cdot(y_{iM}-\mathbf{z}_{iM}^{'}\tilde{\boldsymbol{\delta}}_M) \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{i1}\cdot y_{i1} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{iM}\cdot y_{iM} \end{bmatrix} - \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{i1}\mathbf{z}_{i1}^{'}\tilde{\boldsymbol{\delta}}_1 \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{iM}\mathbf{z}_{iM}^{'}\tilde{\boldsymbol{\delta}}_M \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{i1}\cdot y_{i1} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{iM}\cdot y_{iM} \end{bmatrix} - \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{i1}\mathbf{z}_{i1}^{'} & & \\ & \ddots & \\ & & \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{iM}\mathbf{z}_{iM}^{'} \end{bmatrix} \begin{bmatrix} \tilde{\delta}_1 \\ \vdots \\ \tilde{\delta}_M \end{bmatrix} \\ &\equiv \mathbf{s}_{xy} - \mathbf{S}_{xz}\tilde{\boldsymbol{\delta}} \end{aligned}$$

$$\mathbf{S} = E(\mathbf{g}_i\mathbf{g}_i^{'}) = \begin{bmatrix} E(\varepsilon_{i1}\varepsilon_{i1}\mathbf{x}_{i1}\mathbf{x}_{i1}^{'}) & \dots & E(\varepsilon_{i1}\varepsilon_{iM}\mathbf{x}_{i1}\mathbf{x}_{iM}^{'}) \\ \vdots & & \vdots \\ E(\varepsilon_{iM}\varepsilon_{i1}\mathbf{x}_{iM}\mathbf{x}_{i1}^{'}) & \dots & E(\varepsilon_{iM}\varepsilon_{iM}\mathbf{x}_{iM}\mathbf{x}_{iM}^{'}) \end{bmatrix}$$

$$\widehat{\mathbf{W}} = \begin{bmatrix} \widehat{\mathbf{W}}_{11} & \widehat{\mathbf{W}}_{12} & \dots & \widehat{\mathbf{W}}_{1M} \\ \widehat{\mathbf{W}}'_{12} & \widehat{\mathbf{W}}_{22} & \dots & \widehat{\mathbf{W}}_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\mathbf{W}}'_{1M} & \widehat{\mathbf{W}}'_{2M} & \dots & \widehat{\mathbf{W}}_{MM} \end{bmatrix}$$

MEGMM estimator:

$$\begin{aligned} \tilde{\delta}(\hat{W}) &= \arg\min_{\tilde{\delta}} J(\tilde{\delta}, \hat{W}) = \arg\min_{\tilde{\delta}} n g_n(\tilde{\delta})' \hat{W} g_n(\tilde{\delta}) \\ &= \arg\min_{\tilde{\delta}} n(s_{xy} - S_{xz}\tilde{\delta})' \hat{W} (s_{xy} - S_{xz}\tilde{\delta}) \end{aligned}$$

FOC:

$$\tilde{\delta}(\hat{W}) = (S'_{xz}\hat{W}S_{xz})^{-1} S'_{xz}\hat{W}s_{xy} =$$

$$\begin{pmatrix} \mathbf{S}'_{x_1z_1}\hat{\mathbf{W}}_{11}\mathbf{S}_{x_1z_1} & \mathbf{S}'_{x_1z_1}\hat{\mathbf{W}}_{12}\mathbf{S}_{x_2z_2} & \dots & \mathbf{S}'_{x_1z_1}\hat{\mathbf{W}}_{1M}\mathbf{S}_{x_Mz_M} \\ \mathbf{S}'_{x_2z_2}\hat{\mathbf{W}}'_{12}\mathbf{S}_{x_1z_1} & \mathbf{S}'_{x_2z_2}\hat{\mathbf{W}}_{22}\mathbf{S}_{x_2z_2} & \dots & \mathbf{S}'_{x_2z_2}\hat{\mathbf{W}}_{2M}\mathbf{S}_{x_Mz_M} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}'_{x_Mz_M}\hat{\mathbf{W}}'_{1M}\mathbf{S}_{x_1z_1} & \mathbf{S}'_{x_Mz_M}\hat{\mathbf{W}}'_{2M}\mathbf{S}_{x_2z_2} & \dots & \mathbf{S}'_{x_Mz_M}\hat{\mathbf{W}}_{MM}\mathbf{S}_{x_Mz_M} \end{pmatrix}^{-1} \times \begin{pmatrix} \mathbf{S}'_{x_1z_1}\hat{\mathbf{W}}_{1M}\mathbf{s}_{x_My_M} \\ \mathbf{S}'_{x_2z_2}\hat{\mathbf{W}}_{2M}\mathbf{s}_{x_My_M} \\ \vdots \\ \mathbf{S}'_{x_Mz_M}\hat{\mathbf{W}}_{MM}\mathbf{s}_{x_My_M} \end{pmatrix}$$

SEGMM versus MEGMM

SEGMM is a special case of MEGMM when:

$$\hat{\mathbf{W}} = \text{diag}(\hat{\mathbf{W}}_{11}, \dots, \hat{\mathbf{W}}_{MM})$$

Efficient MEGMM is equivalent to Efficient SEGMM when:

1. Each equation is just identified.
2. At least one equation is over-identified but S is block diagonal.

$$E[g_{im}g'_{ih}] = E[\varepsilon_{im}\varepsilon_{ih}\, x_{im}x'_{ih}] = 0 \quad \text{for all } m \neq h$$

$$\hat{\boldsymbol{\delta}}(\widehat{\mathbf{W}}) = \begin{bmatrix} \hat{\boldsymbol{\delta}}_1(\widehat{\mathbf{W}}) \\ \vdots \\ \hat{\boldsymbol{\delta}}_M(\widehat{\mathbf{W}}) \end{bmatrix} = \left(\begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_{i1}\mathbf{x}_{i1}' & & \\ & \ddots & \\ & & \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_{iM}\mathbf{x}_{iM}' \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{W}}_{11} \cdots \widehat{\mathbf{W}}_{1M} \\ \vdots \quad \vdots \\ \widehat{\mathbf{W}}_{M1} \cdots \widehat{\mathbf{W}}_{MM} \end{bmatrix} \right.$$

$$\left. \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{i1}\mathbf{z}_{i1}' & & \\ & \ddots & \\ & & \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{iM}\mathbf{z}_{iM}' \end{bmatrix} \right)^{-1}$$

$$\begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_{i1}\mathbf{x}_{i1}' & & \\ & \ddots & \\ & & \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_{iM}\mathbf{x}_{iM}' \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{W}}_{11} \cdots \widehat{\mathbf{W}}_{1M} \\ \vdots \quad \vdots \\ \widehat{\mathbf{W}}_{M1} \cdots \widehat{\mathbf{W}}_{MM} \end{bmatrix} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{i1} \cdot y_{i1} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{iM} \cdot y_{iM} \end{bmatrix}$$

$$= \begin{bmatrix} \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_{i1}\mathbf{x}_{i1}'\right)\widehat{\mathbf{W}}_{11}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{i1}\mathbf{z}_{i1}'\right) & \cdots & \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_{i1}\mathbf{x}_{i1}'\right)\widehat{\mathbf{W}}_{1M}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{iM}\mathbf{z}_{iM}'\right) \\ \vdots & & \vdots \\ \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_{iM}\mathbf{x}_{iM}'\right)\widehat{\mathbf{W}}_{M1}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{i1}\mathbf{z}_{i1}'\right) & \cdots & \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_{iM}\mathbf{x}_{iM}'\right)\widehat{\mathbf{W}}_{MM}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{iM}\mathbf{z}_{iM}'\right) \end{bmatrix}^{-1}$$

$$\begin{bmatrix} \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_{i1}\mathbf{x}_{i1}'\right)\widehat{\mathbf{W}}_{11}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{i1}y_{i1}'\right) + \cdots + \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_{i1}\mathbf{x}_{i1}'\right)\widehat{\mathbf{W}}_{1M}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{iM}y_{iM}'\right) \\ \vdots \\ \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_{iM}\mathbf{x}_{iM}'\right)\widehat{\mathbf{W}}_{M1}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{i1}\, y_{i1}\right) + \cdots + \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_{iM}\mathbf{x}_{iM}'\right)\widehat{\mathbf{W}}_{MM}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{iM}\, y_{iM}\right) \end{bmatrix}. \tag{4.2.6}$$

To go from the second-to-last to the last line, use formulas (A.6) and (A.7) of the appendix on partitioned matrices. If you find the operation too much, just set $M = 2$.

Table 4.1: Multiple-Equation GMM in the Single-Equation Format

Sample Analogue of Orthogonality Conditions: $\mathbf{g}_n(\tilde{\delta}) = \mathbf{s}_{xy} - \mathbf{S}_{xz}\tilde{\delta} = \mathbf{0}$

GMM Estimator: $\hat{\delta}(\widehat{\mathbf{W}}) = (\mathbf{S}'_{xz}\widehat{\mathbf{W}}\mathbf{S}_{xz})^{-1}\mathbf{S}'_{xz}\widehat{\mathbf{W}}\mathbf{s}_{xy}$

Its Sampling Error: $\hat{\delta}(\widehat{\mathbf{W}}) - \delta = (\mathbf{S}'_{xz}\widehat{\mathbf{W}}\mathbf{S}_{xz})^{-1}\mathbf{S}'_{xz}\widehat{\mathbf{W}}\bar{\mathbf{g}}$

Asymptotic Variance of Optimal GMM: $\text{Avar}(\hat{\delta}(\widehat{\mathbf{S}}^{-1})) = (\mathbf{\Sigma}'_{xz}\mathbf{S}^{-1}\mathbf{\Sigma}_{xz})^{-1}$

Its Estimator: $\widehat{\text{Avar}(\hat{\delta}(\widehat{\mathbf{S}}^{-1}))} = (\mathbf{S}'_{xz}\widehat{\mathbf{S}}^{-1}\mathbf{S}_{xz})^{-1}$

J Statistic: $J(\hat{\delta}(\widehat{\mathbf{S}}^{-1}), \widehat{\mathbf{S}}^{-1}) \equiv n \cdot \mathbf{g}_n(\hat{\delta}(\widehat{\mathbf{S}}^{-1}))'\widehat{\mathbf{S}}^{-1}\mathbf{g}_n(\hat{\delta}(\widehat{\mathbf{S}}^{-1}))$

	Single-Equation GMM applied to the equation in question	Multiple-Equation GMM
$\mathbf{g}_i$	$\mathbf{x}_i \cdot \varepsilon_i$	(4.1.4)
δ	δ	(4.1.6)
$\mathbf{s}_{xy}$	$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \cdot y_i$	(4.2.2)
$\mathbf{S}_{xz}$	$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{z}'_i$	(4.2.2)
Size of $\mathbf{W}$	$K \times K$	$\sum_m K_m \times \sum_m K_m$
$\mathbf{\Sigma}_{xz}$	$\text{E}(\mathbf{x}_i\mathbf{z}'_i)$	(4.1.9)
$\mathbf{S}$ ($\equiv \text{Avar}(\bar{\mathbf{g}})$)	$\text{E}(\varepsilon_i^2\mathbf{x}_i\mathbf{x}'_i)$	(4.1.11)
$\widehat{\mathbf{S}}$	$\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2\mathbf{x}_i\mathbf{x}'_i$	(4.3.2)
Estimator consistent under which assumptions?	3.1–3.4	4.1–4.4
Estimator asymptotic normal under which assumptions?	3.1–3.5	4.1–4.5
$\widehat{\mathbf{S}} \to_{\text{p}} \mathbf{S}$ under which assumptions?	3.1, 3.2, 3.6, $\text{E}(\mathbf{g}_i\mathbf{g}'_i)$ finite	4.1, 4.2, 4.6, $\text{E}(\mathbf{g}_i\mathbf{g}'_i)$ finite
d.f. of J	$K - L$	$\sum_m (K_m - L_m)$

Special Cases of MEGMM: FIVE, 3SLS and SUR

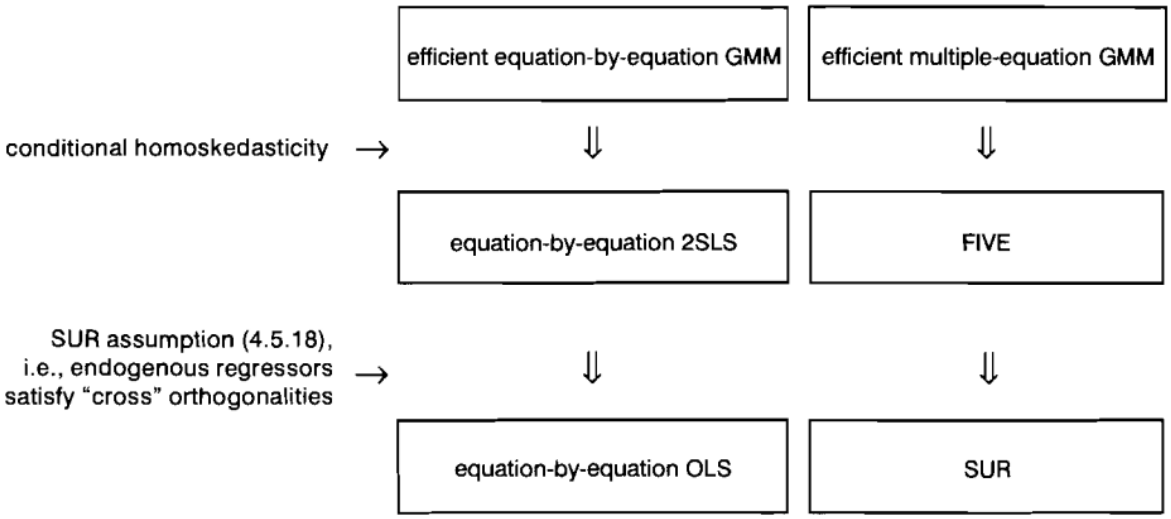


Figure 4.1: OLS and GMM

Full-Information Instrumental Variables Efficient (FIVE)

Under Ass 4.1-4.5 and 4.7,

$$E(\varepsilon_{im}\varepsilon_{ih}|\mathbf{x}_{im},\mathbf{x}_{ih}) = \sigma_{mh}$$

$$\mathbf{S} = E[\mathbf{x}_{im}\mathbf{x}_{ih}^{'}\varepsilon_{im}\varepsilon_{ih}] = \sigma_{mh}E[\mathbf{x}_{im}\mathbf{x}_{ih}^{'}] = \begin{bmatrix} \sigma_{11}E(\mathbf{x}_{i1}\mathbf{x}_{i1}^{'}) & \dots & \sigma_{1M}E(\mathbf{x}_{i1}\mathbf{x}_{iM}^{'}) \\ \vdots & \ddots & \vdots \\ \sigma_{M1}E(\mathbf{x}_{iM}\mathbf{x}_{i1}^{'}) & \dots & \sigma_{MM}E(\mathbf{x}_{iM}\mathbf{x}_{iM}^{'}) \end{bmatrix}$$

$\hat{\sigma}_{mh}^2 \xrightarrow{p} \sigma_{mh}^2$($E(z_{im}z_{ih}')$ exists and finite), $\frac{1}{n}\sum x_{im}x_{ih}' \xrightarrow{p} E(x_{im}x_{ih}')$, $\hat{S} \xrightarrow{p} S$

Large-Sample properties:

1. $\tilde{\delta}_{FIVE} = \tilde{\delta}(\hat{S}^{-1})$ is consistent, asymptotically normal, and efficient with $Avar(\tilde{\delta}(\hat{S}^{-1}))$.

$$Avar(\tilde{\delta}_{FIVE}) = (\Sigma_{xz}'S^{-1}\Sigma_{xz})^{-1}$$

2. $\widehat{Avar}(\tilde{\delta}_{FIVE})$ is consistent for $Avar(\tilde{\delta}_{FIVE})$:

$$\widehat{Avar}(\tilde{\delta}_{FIVE}) = (S_{xz}'\hat{S}^{-1}S_{xz})^{-1}$$

3. Sargan's Statistic:

$$J(\tilde{\delta}_{FIVE},\hat{S}^{-1}) \equiv n \cdot g_n(\tilde{\delta}_{FIVE})'\hat{S}^{-1}g_n(\tilde{\delta}_{FIVE}) \xrightarrow{d} \chi^2(\sum_m (K_m - L_m))$$

Three-Stage Least Squares (3SLS)

Under Ass 4.1-4.5 and 4.7 + same instruments for all equations:
$\mathbf{x}_{i1} = \mathbf{x}_{i2} = \ldots = \mathbf{x}_{im} = \mathbf{x}_i$

moment conditions:

$$\mathbf{g}_i = \begin{bmatrix} \mathbf{x}_i \varepsilon_{i1} \\ \vdots \\ \mathbf{x}_i \varepsilon_{iM} \end{bmatrix} = \boldsymbol{\varepsilon}_i \otimes \mathbf{x}_i, \quad \text{with} \quad \boldsymbol{\varepsilon}_i \equiv \begin{bmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iM} \end{bmatrix}$$

Efficient Weight matrix:

$$\mathbf{S}_{3SLS} = \begin{bmatrix} \sigma_{11} E[\mathbf{x}_i \mathbf{x}_i^{'}] & \sigma_{12} E[\mathbf{x}_i \mathbf{x}_i^{'}] & \ldots & \sigma_{1M} E[\mathbf{x}_i \mathbf{x}_i^{'}] \\ \sigma_{12} E[\mathbf{x}_i \mathbf{x}_i^{'}] & \sigma_{22} E[\mathbf{x}_i \mathbf{x}_i^{'}] & \ldots & \sigma_{2M} E[\mathbf{x}_i \mathbf{x}_i^{'}] \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1M} E[\mathbf{x}_i \mathbf{x}_i^{'}] & \sigma_{2M} E[\mathbf{x}_i \mathbf{x}_i^{'}] & \ldots & \sigma_{MM} E[\mathbf{x}_i \mathbf{x}_i^{'}] \end{bmatrix} = \boldsymbol{\Sigma} \otimes E[\mathbf{x}_i \mathbf{x}_i^{'}] \quad ,$$

$$\text{with} \quad \boldsymbol{\Sigma} = E[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i^{'}] = \begin{bmatrix} \boldsymbol{\sigma}_{11} & \boldsymbol{\sigma}_{12} & \ldots & \boldsymbol{\sigma}_{1M} \\ \boldsymbol{\sigma}_{12} & \boldsymbol{\sigma}_{22} & \ldots & \boldsymbol{\sigma}_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\sigma}_{1M} & \boldsymbol{\sigma}_{2M} & \ldots & \boldsymbol{\sigma}_{MM} \end{bmatrix}$$

$$\widehat{\boldsymbol{\Sigma}} \equiv \begin{bmatrix} \hat{\boldsymbol{\sigma}}_{11} & \ldots & \hat{\boldsymbol{\sigma}}_{1M} \\ \vdots & & \vdots \\ \hat{\boldsymbol{\sigma}}_{M1} & \ldots & \hat{\boldsymbol{\sigma}}_{MM} \end{bmatrix} = \frac{1}{n} \sum_{i=1}^{n} \hat{\boldsymbol{\varepsilon}}_i \hat{\boldsymbol{\varepsilon}}_i^{'}, \quad S_{xx} = \frac{1}{n} \mathbf{X}' \mathbf{X}$$

Large-Sample properties:

1. $\tilde{\delta}_{3SLS}$ is consistent, asymptotically normal, and efficient with $Avar(\tilde{\delta}_{3SLS})$.
2. $\widehat{Avar}(\tilde{\delta}_{3SLS})$ is consistent for $Avar(\tilde{\delta}_{3SLS})$.
3. Sargan's Statistic:

$$J(\tilde{\delta}_{\text{3SLS}}, \hat{S}^{-1}) \equiv n \cdot g_n(\tilde{\delta}_{\text{3SLS}})' \hat{S}^{-1} g_n(\tilde{\delta}_{3SLS}) \xrightarrow{d} \chi^2(MK - \sum_m L_m)$$

Seemingly Unrelated Regression (SUR)

Under Ass 4.1-4.5 and 4.7 + same instruments for all equations + the predetermined regressors satisfy "cross" orthogonalities: $E[\mathbf{z}_{im} \varepsilon_{ih}] = 0$ (without endogeneity)

Efficient Weight Matrix:

$$\mathbf{S}_{SUR} = \begin{bmatrix} \sigma_{11}E[\mathbf{z}_i\mathbf{z}_i^{'}] & \sigma_{12}E[\mathbf{z}_i\mathbf{z}_i^{'}] & \dots & \sigma_{1M}E[\mathbf{z}_i\mathbf{z}_i^{'}] \\ \sigma_{12}E[\mathbf{z}_i\mathbf{z}_i^{'}] & \sigma_{22}E[\mathbf{z}_i\mathbf{z}_i^{'}] & \dots & \sigma_{2M}E[\mathbf{z}_i\mathbf{z}_i^{'}] \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1M}E[\mathbf{z}_i\mathbf{z}_i^{'}] & \sigma_{2M}E[\mathbf{z}_i\mathbf{z}_i^{'}] & \dots & \sigma_{MM}E[\mathbf{z}_i\mathbf{z}_i^{'}] \end{bmatrix} = \mathbf{\Sigma} \otimes E[\mathbf{z}_i\mathbf{z}_i^{'}]$$

Large-Sample properties:

1. $\tilde{\delta}_{SUR}$ is consistent, asymptotically normal, and efficient with $Avar(\tilde{\delta}_{SUR})$.
2. $\widehat{Avar}(\tilde{\delta}_{SUR})$ is consistent for $Avar(\tilde{\delta}_{SUR})$.
3. Sargan's Statistic:

$$J(\tilde{\delta}_{SUR}, \hat{S}^{-1}) \equiv n \cdot g_n(\tilde{\delta}_{SUR})'\hat{S}^{-1}g_n(\tilde{\delta}_{SUR}) \xrightarrow{d} \chi^2(MK - \sum_m L_m)$$

SUR versus OLS

the same as SEGMM versus MEGMM:

1. Each equation is just identified.
2. At least one equation is over-identified. Need assumption that

$$\sigma_{mh}E(x_i x_i') = 0 \text{ for all m} \neq \text{h}$$

Table 4.2: Relationship between Multiple-Equation Estimators

	Multiple-equation GMM	FIVE	3SLS	SUR	Multivariate regression
The model	Assumptions 4.1–4.6	Assumptions 4.1–4.5, Assumption 4.7 $\mathrm{E}(\mathbf{z}_{im}\mathbf{z}_{ih}')$ finite	Assumptions 4.1–4.5, Assumption 4.7 $\mathrm{E}(\mathbf{z}_{im}\mathbf{z}_{ih}')$ finite $\mathbf{x}_{im} = \mathbf{x}_i$ for all m	Assumptions 4.1–4.5, Assumption 4.7 $\mathbf{x}_{im} = \mathbf{x}_i$ for all m $\mathbf{x}_i$ = union of $\mathbf{z}_{i1}, \ldots, \mathbf{z}_{iM}$	Assumptions 4.1–4.5, Assumption 4.7 $\mathbf{x}_{im} = \mathbf{x}_i$ for all m $\mathbf{z}_{im} = \mathbf{x}_i$ for all m
$\mathbf{S}$ ($\equiv$ Avar($\bar{\mathbf{g}}$))	(4.1.11)	(4.5.2)	$\mathbf{\Sigma} \otimes \mathrm{E}(\mathbf{x}_i\mathbf{x}_i')$	$\mathbf{\Sigma} \otimes \mathrm{E}(\mathbf{x}_i\mathbf{x}_i')$	irrelevant
$\widehat{\mathbf{S}}$	(4.3.2)	(4.5.3)	$\widehat{\mathbf{\Sigma}} \otimes (n^{-1}\sum_i \mathbf{x}_i\mathbf{x}_i')$ $\widehat{\mathbf{\Sigma}}$ from 2SLS residuals	$\widehat{\mathbf{\Sigma}} \otimes (n^{-1}\sum_i \mathbf{x}_i\mathbf{x}_i')$ $\widehat{\mathbf{\Sigma}}$ from OLS residuals	irrelevant
$\hat{\boldsymbol{\delta}}(\widehat{\mathbf{S}}^{-1})$	no simplification	no simplification	(4.5.12) with (4.5.13), (4.5.14)	(4.5.12) with (4.5.13′), (4.5.14′)	equation-by-equation OLS
Avar($\hat{\boldsymbol{\delta}}(\widehat{\mathbf{S}}^{-1})$)	(4.3.3)	(4.3.3)	(4.5.15) with (4.5.16)	(4.5.15) with (4.5.16′)	OLS formula
$\widehat{\text{Avar}}(\hat{\boldsymbol{\delta}}(\widehat{\mathbf{S}}^{-1}))$	(4.3.4)	(4.3.4)	(4.5.17) with (4.5.13)	(4.5.17) with (4.5.13′)	OLS formula

Three-Stage Least Squares (3SLS)

Three-Stage Least Squares is a **systems estimation method** for simultaneous equation models. Unlike single-equation methods like 2SLS, it estimates **all equations simultaneously**, improving efficiency when error terms across equations are correlated.

2SLS vs. 3SLS

	2SLS	3SLS
Estimation	Equation-by-equation	All equations
Information Used	Within-equation information only	Full system information
Cross-equation Error Correlations	Ignored	Considered
Efficiency	Lower	Higher

If error terms across equations are **uncorrelated**, 3SLS reduces to 2SLS. In practice, economic relationships often create correlated errors due to omitted variables, making 3SLS more efficient.

Estimation

Prerequisites

1. **No autocorrelation** within each equation's error term.
2. **Contemporaneous correlation** between error terms of different equations.
3. The system must be **overidentified** (verified via order and rank conditions).

Model Setup

$$y_{im} = z'_{im}\delta + \varepsilon_{im}$$

$$y = (Y\gamma + X\beta) + \varepsilon = Z\delta + \varepsilon$$

Stage 1 (Same as 2SLS)

reduced form:

$$Z = X\pi + v$$

Stage 2 (Same as 2SLS)

structural model:

$$y = \hat{Z}\delta + u$$

Stage 3

1. Estimate the **cross-equation covariance matrix** Σ using residuals from Stage 2:

$$\hat{\sigma}_{ij} = \frac{\hat{u}_i' \hat{u}_j}{n}$$

$$\hat{\Sigma} = \begin{bmatrix} \widehat{\sigma}_1^2 & \widehat{\sigma}_{12} & \cdots & \widehat{\sigma}_{1m} \\ \widehat{\sigma}_{21} & \widehat{\sigma}_2^2 & \cdots & \widehat{\sigma}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\sigma}_{m1} & \widehat{\sigma}_{m2} & \cdots & \widehat{\sigma}_m^2 \end{bmatrix}$$

2. Apply **Generalized Least Squares (GLS)** to the entire system. The **3SLS estimator** is:

$$\hat{\delta}_{3SLS} = \left[\hat{Z}' \left(\hat{\Sigma}^{-1} \otimes I\right) \hat{Z}\right]^{-1} \hat{Z}' \left(\hat{\Sigma}^{-1} \otimes I\right) y$$

Reference:

- 3SLS: Three-Stage Least Squares - SPUR ECONOMICS
- 陈强编著.高级计量经济学及Stata应用.第2版.高等教育出版社.2014

Binary choice model

Logit & Probit model

$$G(z) = \frac{e^z}{1+e^z}$$

$$g(z) = \frac{e^z}{(1+e^z)^2}$$

$$\Phi(z) \equiv \int_{-\infty}^{z} \phi(v)dv$$

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Coefficient estimates and standard errors in the logit model are roughly $\pi/\sqrt{3} = 1.8$ bigger than in the probit model.

Odds ratio= $\frac{P(y=1|x=1)}{P(y=1|x=0)}$

Latent variable

y_i^* is unobserved, it is referred to as a latent variable.

$$y_i^* = x_i'\beta + \varepsilon_i. \tag{7.8}$$

$$\begin{aligned} y_i &= 1 \quad \text{if } y_i^* > 0 \\ &= 0 \quad \text{if } y_i^* \le 0, \end{aligned} \tag{7.10}$$

$$P\{y_i = 1\} = P\{y_i^* > 0\} = P\{x_i'\beta + \varepsilon_i > 0\} = P\{-\varepsilon_i \le x_i'\beta\} = F(x_i'\beta), \tag{7.9}$$

Estimation:

$$L(\beta) = \prod_{i=1}^{N} P\{y_i = 1|x_i;\beta\}^{y_i} P\{y_i = 0|x_i;\beta\}^{1-y_i}, \tag{7.11}$$

$$\log L(\beta) = \sum_{i=1}^{N} y_i \log F(x_i'\beta) + \sum_{i=1}^{N} (1-y_i)\log(1-F(x_i'\beta)). \tag{7.12}$$

$$\frac{\partial \log L(\beta)}{\partial \beta} = \sum_{i=1}^{N} \left[\frac{y_i - F(x_i'\beta)}{F(x_i'\beta)(1-F(x_i'\beta))} f(x_i'\beta) \right] x_i = 0, \tag{7.13}$$

Goodness-of-fit:
Goodness-of-fit measures are based on comparison with a model that contains only a constant as explanatory variable. Let logL1 denote the maximum loglikelihood value of the model of interest and let logL0 denote the maximum value of the loglikelihood function when all parameters, except the intercept, are set to zero. Clearly, logL1 ≥

logL0. The larger the difference between the two loglikelihood values, the more the extended model adds to the very restrictive model.

1. $$pseudoR^2 = 1 - \frac{1}{1 + 2(\log L_1 - \log L_0)/N}, \tag{7.17}$$
2. $$McFaddenR^2 = 1 - \log L_1 / \log L_0, \tag{7.18}$$

Table 7.1 Cross-tabulation of actual and predicted outcomes

		$\hat{y}_i$		
		0	1	Total
y_i	0	n_{00}	n_{01}	N_0
	1	n_{10}	n_{11}	N_1
	Total	n_0	n_1	N

$$\hat{p} = N_1/N,$$

the proportion of incorrect predictions:

$$wr_1 = \frac{n_{01} + n_{10}}{N},$$

$$\begin{aligned} wr_0 &= 1 - \hat{p} \quad \text{if } \hat{p} > 0.5, \\ &= \hat{p} \quad\quad \text{if } \hat{p} \leq 0.5. \end{aligned}$$

3. $$R_p^2 = 1 - \frac{wr_1}{wr_0}. \tag{7.21}$$

Illustration: the Impact of Unemployment Benefits on Recipiency

Table 7.2 Binary choice models for applying for unemployment benefits (blue collar workers)

	LPM		Logit		Probit	
Variable	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.
constant	−0.077	(0.122)	−2.800	(0.604)	−1.700	(0.363)
replacement rate	0.629	(0.384)	3.068	(1.868)	1.863	(1.127)
replacement rate^2	−1.019	(0.481)	−4.891	(2.334)	−2.980	(1.411)
age	0.0157	(0.0047)	0.068	(0.024)	0.042	(0.014)
age^2/10	−0.0015	(0.0006)	−0.0060	(0.0030)	−0.0038	(0.0018)
tenure	0.0057	(0.0012)	0.0312	(0.0066)	0.0177	(0.0038)
slack work	0.128	(0.014)	0.625	(0.071)	0.375	(0.042)
abolished position	−0.0065	(0.0248)	−0.0362	(0.1178)	−0.0223	(0.0718)
seasonal work	0.058	(0.036)	0.271	(0.171)	0.161	(0.104)
head of household	−0.044	(0.017)	−0.211	(0.081)	−0.125	(0.049)
married	0.049	(0.016)	0.242	(0.079)	0.145	(0.048)
children	−0.031	(0.017)	−0.158	(0.086)	−0.097	(0.052)
young children	0.043	(0.020)	0.206	(0.097)	0.124	(0.059)
live in SMSA	−0.035	(0.014)	−0.170	(0.070)	−0.100	(0.042)
non-white	0.017	(0.019)	0.074	(0.093)	0.052	(0.056)
year of displacement	−0.013	(0.008)	−0.064	(0.015)	−0.038	(0.009)
>12 years of school	−0.014	(0.016)	−0.065	(0.082)	−0.042	(0.050)
male	−0.036	(0.018)	−0.180	(0.088)	−0.107	(0.053)
state max. benefits	0.0012	(0.0002)	0.0060	(0.0010)	0.0036	(0.0006)
state unempl. rate	0.018	(0.003)	0.096	(0.016)	0.057	(0.009)
Loglikelihood			−2873.197		−2874.071	
Pseudo R^2			0.066		0.066	
McFadden R^2			0.057		0.057	
R^2_p	0.035		0.046		0.045	

The replacement rate, defined as the ratio of weekly UI(unemployment insurance) benefits to previous weekly earnings.

The LPM is estimated by OLS, the logit and probit model are both estimated by maximum likelihood. The estimates of β obtained from the logit model are roughly a factor π/√3 larger than those obtained from the probit model.

The replacement rate has an insignificant positive coefficient, while its square is significantly negative. For the probit model, we can derive that the estimated marginal effect of a change in the replacement rate (rr) equals **the value of the normal density function** multiplied by 1.863 − 2 × 2.980rr.

The dummy variable which indicates whether the job was lost because of slack work is highly significant in all specifications, which is not surprising given that these workers typically will find it hard to get a new job.

The higher the state unemployment rate and the higher the maximum benefit level, the more likely it is that individuals apply for benefits.

The ceteris paribus effect of being married is estimated to be positive, while, somewhat surprisingly, being head of the household has a negative effect on the probability of take-up.

Table 7.3 Cross-tabulation of actual and predicted outcomes (logit model)

		$\hat{y}_i$		
		0	1	Total
y_i	0	242	1300	1542
	1	171	3164	3335
	Total	413	4464	4877

$$R_p^2 = 1 - \frac{171 + 1300}{1542},$$

$\hat{p} = 3335/4877$

$$\log L_0 = 3335 \log \frac{3335}{4877} + 1542 \log \frac{1542}{4877} = -3046.187,$$

which allows us to compute the pseudo and McFadden R2 measures.
$p_{00} + p_{11} =$

$$\frac{242}{1542} + \frac{3164}{3335} = 1.106,$$

Multi-response Models

Ordered Response Models

$$y_i^* = x_i'\beta + \varepsilon_i \tag{7.28}$$
$$y_i = j \quad \text{if } \gamma_{j-1} < y_i^* \le \gamma_j, \tag{7.29}$$

$$y_i^* = x_i'\beta + \varepsilon_i \tag{7.30}$$
$$\begin{aligned} y_i &= 1 \quad \text{if } y_i^* \le 0, \\ &= 2 \quad \text{if } 0 < y_i^* \le \gamma, \\ &= 3 \quad \text{if } y_i^* > \gamma, \end{aligned} \tag{7.31}$$

Assuming that εi is i.i.d. standard normal results in the ordered probit model. The logistic distribution gives the ordered logit model. For M = 2 we are back at the binary choice model.

$$P\{y_i = 1|x_i\} = P\{y_i^* \le 0|x_i\} = \Phi(-x_i'\beta),$$
$$P\{y_i = 3|x_i\} = P\{y_i^* > \gamma|x_i\} = 1 - \Phi(\gamma - x_i'\beta)$$

and

$$P\{y_i = 2|x_i\} = \Phi(\gamma - x_i'\beta) - \Phi(-x_i'\beta),$$

Normalization:

$$P\{y_i = 1|x_i\} = P\{\beta_1 + x_i'\beta + \varepsilon_i \le \gamma_1|x_i\} = \Phi\left(\frac{\gamma_1 - \beta_1}{\sigma} - x_i'\left(\frac{\beta}{\sigma}\right)\right),$$

Illustration: Willingness to Pay for Natural Areas

Table 7.4 Ordered probit model for willingness to pay

	I: intercept only		II: with characteristics	
Variable	Estimate	s.e.	Estimate	s.e.
constant	18.74	(2.77)	30.55	(8.59)
age class	–		−6.93	(1.64)
female	–		−5.88	(5.07)
income class	–		4.86	(1.87)
$\hat{\sigma}$	38.61	(2.11)	36.47	(1.89)
Loglikelihood	−409.00		−391.25	
Normality test (χ_2^2)	6.326	($p = 0.042$)	2.419	($p = 0.298$)

Multinomial Models

$$\begin{aligned} P\{y_i = j\} &= P\{U_{ij} = \max\{U_{i1}, \ldots, U_{iM}\}\} \\ &= P\left\{\mu_{ij} + \varepsilon_{ij} > \max_{k=1,\ldots,J,k\neq j}\{\mu_{ik} + \varepsilon_{ik}\}\right\}. \end{aligned} \tag{7.35}$$

we assume that all εij are mutually independent with a so-called log Weibull distribution (also known as a Type I extreme value distribution).

$$F(t) = \exp\{-e^{-t}\}, \tag{7.36}$$

$$P\{y_i = j\} = \frac{\exp\{\mu_{ij}\}}{\exp\{\mu_{i1}\} + \exp\{\mu_{i2}\} + \cdots + \exp\{\mu_{iM}\}}. \tag{7.37}$$

multinomial logit model:

$$P\{y_i = j\} = \frac{\exp\{x_{ij}'\beta\}}{1 + \exp\{x_{i2}'\beta\} + \cdots + \exp\{x_{iM}'\beta\}}, \quad j = 1, 2, \ldots, M. \tag{7.38}$$

independence of irrelevant alternatives (IIA)

The probability ratio (or odds ratio) is given by:

$$\frac{P(y_i = j)}{P(y_i = k)} = \frac{exp(x_{ij}'\beta)}{exp(x_{ik}'\beta)}$$

Models for Count Data

Poisson regression model

$$P\{y_i = y|x_i\} = \frac{\exp\{-\lambda_i\}\lambda_i^y}{y!}, \quad y = 0, 1, 2, \ldots, \tag{7.42}$$

estimation:

$$\begin{aligned}\log L(\beta) &= \sum_{i=1}^{N}[-\lambda_i + y_i \log \lambda_i - \log y_i!] \qquad (7.44)\\ &= \sum_{i=1}^{N}[-\exp\{x_i'\beta\} + y_i(x_i'\beta) - \log y_i!].\end{aligned}$$

equidispersion:

$$V\{y_i|x_i\} = \exp\{x_i'\beta\}. \tag{7.43}$$

overdispersion:

$$V\{y_i|x_i\} = E\{\varepsilon_i^2|x_i\} > \exp\{x_i'\beta\}$$

NegBin I model:

$$V\{y_i|x_i\} = (1 + \delta^2)\exp\{x_i'\beta\} \tag{7.49}$$

NegBin II model:

$$V\{y_i|x_i\} = (1 + \alpha^2 \exp\{x_i'\beta\})\exp\{x_i'\beta\}, \tag{7.50}$$

Illustration: Patents and R&D Expenditures

Table 7.5 Estimation results Poisson model, MLE and QMLE

	MLE		QMLE
	Estimate	Standard error	Robust s.e.
constant	−0.8737	0.0659	0.7429
log (*R&D*)	0.8545	0.0084	0.0937
aerospace	−1.4218	0.0956	0.3802
chemistry	0.6363	0.0255	0.2254
computers	0.5953	0.0233	0.3008
machines	0.6890	0.0383	0.4147
vehicles	−1.5297	0.0419	0.2807
Japan	0.2222	0.0275	0.3528
USA	−0.2995	0.0253	0.2736
Loglikelihood	−4950.789		
Pseudo R^2	0.675		
LR test (χ^2_8)	20587.54 ($p = 0.000$)		
Wald test (χ^2_8)			338.9 ($p = 0.000$)

Table 7.6 Estimation results NegBin I and NegBin II model, MLE

	NegBin I (MLE)			NegBin II (MLE)	
	Estimate	Standard error		Estimate	Standard error
constant	0.6899	0.5069		−0.3246	0.4982
log (*R&D*)	0.5784	0.0676		0.8315	0.0766
aerospace	−0.7865	0.3368		−1.4975	0.3772
chemistry	0.7333	0.1852		0.4886	0.2568
computers	0.1450	0.2063		−0.1736	0.2988
machines	0.1559	0.2550		0.0593	0.2793
vehicles	−0.8176	0.2686		−1.5306	0.3739
Japan	0.4005	0.2573		0.2522	0.4264
USA	0.1588	0.1984		−0.5905	0.2788
δ^2	95.2437	14.0069	α^2	1.3009	0.1375
Loglikelihood	−848.195			−819.596	
Pseudo R^2	0.944			0.946	
LR test (χ^2_8)	88.55 ($p = 0.000$)			145.75 ($p = 0.000$)	

The Standard Tobit Model

censored regression model

$$
\begin{aligned}
y_i^* &= x_i'\beta + \varepsilon_i, \quad i = 1, 2, \ldots, N, \\
y_i &= y_i^* \quad \text{if } y_i^* > 0 \\
&= 0 \quad \text{if } y_i^* \le 0,
\end{aligned} \tag{7.60}
$$

$$
\begin{aligned}
P\{y_i = 0\} &= P\{y_i^* \le 0\} = P\{\varepsilon_i \le -x_i'\beta\} \\
&= P\left\{\frac{\varepsilon_i}{\sigma} \le -\frac{x_i'\beta}{\sigma}\right\} = \Phi\left(-\frac{x_i'\beta}{\sigma}\right) = 1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right).
\end{aligned} \tag{7.61}
$$

$$
\frac{\partial P\{y_i = 0\}}{\partial x_{ik}} = -\phi(x_i'\beta/\sigma)\frac{\beta_k}{\sigma}. \tag{7.63}
$$

$$
E\{y_i | y_i > 0\} = x_i'\beta + E\{\varepsilon_i | \varepsilon_i > -x_i'\beta\} = x_i'\beta + \sigma\frac{\phi(x_i'\beta/\sigma)}{\Phi(x_i'\beta/\sigma)}. \tag{7.62}
$$

$$
E\{y_i\} = x_i'\beta\Phi(x_i'\beta/\sigma) + \sigma\phi(x_i'\beta/\sigma). \tag{7.64}
$$

$$
\frac{\partial E\{y_i\}}{\partial x_{ik}} = \beta_k\Phi(x_i'\beta/\sigma). \tag{7.65}
$$

The conditional expectaction of y (for y>0):

$$
\begin{aligned}
\mathrm{E}(y|y > 0, \mathbf{x}) &= \mathbf{x}\beta + \mathrm{E}(u|u > -\mathbf{x}\beta) \\
&= \mathbf{x}\beta + \sigma\mathrm{E}[(u/\sigma)|(u/\sigma) > -\mathbf{x}\boldsymbol{\beta}/\sigma] \\
&= \mathbf{x}\beta + \sigma\phi(\mathbf{x}\beta/\sigma)/\Phi(\mathbf{x}\beta/\sigma) \\
&= \mathbf{x}\boldsymbol{\beta} + \sigma\lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)
\end{aligned}
$$

λ is called the inverse Mill's ratio.

The unconditional expectaction of y (for all y):

$$
\begin{aligned}
E(y|\mathbf{x}) &= P(y > 0|\mathbf{x}) \cdot E(y|y > 0, \mathbf{x}) + P(y = 0|\mathbf{x}) \cdot 0 \\
&= \mathrm{P}(y > 0|\mathbf{x}) \cdot \mathrm{E}(y|y > 0, \mathbf{x}) \\
&= \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma) \cdot \mathrm{E}(y|y > 0, \mathbf{x}) \\
&= \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)[\mathbf{x}\boldsymbol{\beta} + \sigma\lambda(\mathbf{x}\boldsymbol{\beta}/\sigma)] \\
&= \Phi(\mathbf{x}\boldsymbol{\beta}/\sigma)\mathbf{x}\boldsymbol{\beta} + \sigma\phi(\mathbf{x}\boldsymbol{\beta}/\sigma)
\end{aligned}
$$

Estimation

$$
\begin{aligned}
\log L_1(\beta, \sigma^2) &= \sum_{i \in I_0} \log P\{y_i = 0\} + \sum_{i \in I_1} [\log f(y_i | y_i > 0) + \log P\{y_i > 0\}] \\
&= \sum_{i \in I_0} \log P\{y_i = 0\} + \sum_{i \in I_1} \log f(y_i),
\end{aligned} \tag{7.67}
$$

$$\log L_1(\beta, \sigma^2) = \sum_{i \in I_0} \log \left[1 - \Phi \left(\frac{x_i' \beta}{\sigma} \right) \right] + \sum_{i \in I_1} \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2} \frac{(y_i - x_i'\beta)^2}{\sigma^2} \right\} \right]. \quad (7.68)$$

---

truncated regression model

$$y_i^* = x_i'\beta + \varepsilon_i, \quad i = 1, 2, \ldots, N, \quad (7.69)$$
$$y_i = y_i^* \quad \text{if } y_i^* > 0$$
$$(y_i, x_i) \text{ not observed if } y_i^* \leq 0,$$

Estimation:

$$\log L_2(\beta, \sigma^2) = \sum_{i \in I_1} \log f(y_i | y_i > 0) = \sum_{i \in I_1} [\log f(y_i) - \log P\{y_i > 0\}], \quad (7.70)$$

$$\log L_2(\beta, \sigma^2) = \sum_{i \in I_1} \left\{ \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2} \frac{(y_i - x_i'\beta)^2}{\sigma^2} \right\} \right] - \log \Phi \left(\frac{x_i'\beta}{\sigma} \right) \right\}. \quad (7.71)$$

Illustration: Expenditures on Alcohol and Tobacco (Part 1)

Table 7.7 Tobit models for budget shares alcohol and tobacco

	Alcoholic beverages		Tobacco	
Variable	Estimate	s.e.	Estimate	s.e.
constant	−0.1592	(0.0438)	0.5900	(0.0934)
age class	0.0135	(0.0109)	−0.1259	(0.0242)
nadults	0.0292	(0.0169)	0.0154	(0.0380)
nkids ≥ 2	−0.0026	(0.0006)	0.0043	(0.0013)
nkids < 2	−0.0039	(0.0024)	−0.0100	(0.0055)
$\log(x)$	0.0127	(0.0032)	−0.0444	(0.0069)
age $\times \log(x)$	−0.0008	(0.0088)	0.0088	(0.0018)
nadults $\times \log(x)$	−0.0022	(0.0012)	−0.0006	(0.0028)
$\hat{\sigma}$	0.0244	(0.0004)	0.0480	(0.0012)
Loglikelihood	4755.371		758.701	
Wald test (χ_7^2)	117.86	($p = 0.000$)	170.18	($p = 0.000$)

For tabacco, age is an important factor in explaining the budget share. For alcoholic beverages, the number of children and total expenditures are individually significant.

Wald tests for the hypothesis that "all coefficients, except the intercept term, are equal to zero", produce highly significant values for both goods.

The Tobit II Model

The traditional model to describe sample selection problems is the **tobit II model**,[26] also referred to as the **sample selection model**. In this context, it consists of a linear wage equation

$$w_i^* = x_{1i}'\beta_1 + \varepsilon_{1i}, \tag{7.80}$$

where x_{1i} denotes a vector of exogenous characteristics (age, education, gender, ...) and w_i^* denotes person i's wage. The wage w_i^* is not observed for people that are not working (which explains the *). To describe whether a person is working or not a second equation is specified, which is of the binary choice type. That is,

$$h_i^* = x_{2i}'\beta_2 + \varepsilon_{2i}, \tag{7.81}$$

where we have the following observation rule:

$$w_i = w_i^*, h_i = 1 \quad \text{if } h_i^* > 0 \tag{7.82}$$

$$w_i \text{ not observed}, h_i = 0 \quad \text{if } h_i^* \leq 0, \tag{7.83}$$

The conditional expected wage, given that a person *is* working, is given by

$$\begin{aligned} E\{w_i|h_i = 1\} &= x_{1i}'\beta_1 + E\{\varepsilon_{1i}|h_i = 1\} \\ &= x_{1i}'\beta_1 + E\{\varepsilon_{1i}|\varepsilon_{2i} > -x_{2i}'\beta_2\} \\ &= x_{1i}'\beta_1 + \frac{\sigma_{12}}{\sigma_2^2} E\{\varepsilon_{2i}|\varepsilon_{2i} > -x_{2i}'\beta_2\} \\ &= x_{1i}'\beta_1 + \sigma_{12}\frac{\phi(x_{2i}'\beta_2)}{\Phi(x_{2i}'\beta_2)}, \end{aligned} \tag{7.84}$$

the inverse Mill's ratio: Heckman's lambda

Although the tobit II model can be motivated in different ways, we shall more or less follow Gronau (1974) in his reasoning. Assume that the utility maximization problem of the individual (in Gronau's case: housewives) can be characterized by a **reservation wage** w_i^r (the value of time). An individual will work if the actual wage she is offered exceeds this reservation wage. The reservation wage of course depends upon personal characteristics, via the utility function and the budget constraint, so that we write (assume)

$$w_i^r = z_i'\gamma + \eta_i,$$

where z_i is a vector of characteristics and η_i is unobserved. Usually the reservation wage is not observed.

Now assume that the wage a person is offered depends on her personal characteristics (and some job characteristics) as in (7.80), i.e.

$$w_i^* = x_{1i}'\beta_1 + \varepsilon_{1i}.$$

If this wage is below w_i^r individual i is assumed not to work. We can thus write her labour supply decision as

$$\begin{aligned} h_i &= 1 \quad \text{if } w_i^* - w_i^r > 0 \\ &= 0 \quad \text{if } w_i^* - w_i^r \leq 0 \end{aligned}$$

$$h_i^* \equiv w_i^* - w_i^r = x_{1i}'\beta_1 - z_i'\gamma + (\varepsilon_{1i} - \eta_i) = x_{2i}'\beta_2 + \varepsilon_{2i}, \tag{7.85}$$

$$y_i^* = x_{1i}'\beta_1 + \varepsilon_{1i} \tag{7.86}$$

$$h_i^* = x_{2i}'\beta_2 + \varepsilon_{2i} \tag{7.87}$$

$$y_i = y_i^*, h_i = 1 \quad \text{if } h_i^* > 0 \tag{7.88}$$

$$y_i \text{ not observed}, h_i = 0 \quad \text{if } h_i^* \leq 0, \tag{7.89}$$

where

$$\begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \sim NID\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & 1 \end{pmatrix}\right). \tag{7.90}$$

Estimation:

$$\log L_3(\beta, \sigma_1^2, \sigma_{12}) = \sum_{i \in I_0} \log P\{h_i = 0\} + \sum_{i \in I_1} [\log f(y_i | h_i = 1) + \log P\{h_i = 1\}]. \tag{7.91}$$

Heckman's two-step estimation

$$y_i = x_{1i}'\beta_1 + \sigma_{12}\lambda_i + \eta_i, \tag{7.97}$$

where

$$\lambda_i = \frac{\phi(x_{2i}'\beta_2)}{\Phi(x_{2i}'\beta_2)}.$$

- Step 1: Estimate the selection equation(7.87) using Probit model. Obtain the estimate $\hat{\beta}_2$ and compute the Inverse Mills Ratio.
- Step 2: Run OLS on the outcome equation(7.97) using only the selected sample ($h_i = 1$). Obtain estimates $\hat{\beta}_1$ and $\hat{\lambda}$.

In Heckman two-step estimation, since the error from the first step carries over to the second step, its efficiency is lower than the overall estimation of MLE. The advantage of the two-step method lies in its simplicity of operation and weaker assumptions about the distribution. Moreover we can get the probability $P(h_i = 1)$ in the first stage.

Illustration: Expenditures on Alcohol and Tobacco (Part 2)

Table 7.8 Models for budget shares alcohol and tobacco, estimated by OLS using positive observations only

	Alcoholic beverages		Tobacco	
Variable	Estimate	s.e.	Estimate	s.e.
constant	0.0527	(0.0439)	0.4897	(0.0741)
age class	0.0078	(0.0110)	−0.0315	(0.0206)
nadults	−0.0131	(0.0163)	−0.0130	(0.0324)
nkids ≥ 2	−0.0020	(0.0006)	0.0013	(0.0011)
nkids < 2	−0.0024	(0.0023)	−0.0034	(0.0045)
$\log(x)$	−0.0023	(0.0032)	−0.0336	(0.0055)
age $\times \log(x)$	−0.0004	(0.0008)	0.0022	(0.0015)
nadults $\times \log(x)$	0.0008	(0.0012)	0.0011	(0.0023)
	$R^2 = 0.051$	$s = 0.0215$	$R^2 = 0.154$	$s = 0.0291$
	$N = 2258$		$N = 1036$	

Table 7.9 Probit models for abstention of alcohol and tobacco

	Alcoholic beverages		Tobacco	
Variable	Estimate	s.e.	Estimate	s.e.
constant	−15.882	(2.574)	8.244	(2.211)
age	0.6679	(0.6520)	−2.4830	(0.5596)
nadults	2.2554	(1.0250)	0.4852	(0.8717)
nkids ≥ 2	−0.0770	(0.0372)	0.0813	(0.0308)
nkids < 2	−0.1857	(0.1408)	−0.2117	(0.1230)
$\log(x)$	1.2355	(0.1913)	−0.6321	(0.1632)
age $\times \log(x)$	−0.0448	(0.0485)	0.1747	(0.0413)
nadults $\times \log(x)$	−0.1688	(0.0743)	−0.0253	(0.0629)
blue collar	−0.0612	(0.0978)	0.2064	(0.0834)
white collar	0.0506	(0.0847)	0.0215	(0.0694)
Loglikelihood	−1159.865		−1754.886	
Wald test (χ^2_9)	173.18	($p = 0.000$)	108.91	($p = 0.000$)

For alcoholic beverages, total expenditure, the number of adults in the household as well as the number of children older than 2 are statistically significant in explaining abstention.

For tobacco, total expenditure, number of children older than 2, age and being a blue-collar worker are statistically important explanators for abstention.

Table 7.10 Two-step estimation of Engel curves for alcohol and tobacco (tobit II model)

	Alcoholic beverages		Tobacco	
Variable	Estimate	s.e.	Estimate	s.e.
constant	0.0543	(0.0487)	0.4516	(0.0735)
age class	0.0077	(0.0110)	−0.0173	(0.0206)
nadults	−0.0133	(0.0166)	−0.0174	(0.0318)
nkids ≥ 2	−0.0020	(0.0006)	0.0008	(0.0010)
nkids < 2	−0.0024	(0.0023)	−0.0021	(0.0045)
$\log(x)$	−0.0024	(0.0035)	−0.0301	(0.0055)
age $\times \log(x)$	−0.0004	(0.0008)	0.0012	(0.0015)
nadults $\times \log(x)$	0.0008	(0.0012)	−0.0041	(0.0023)
λ	−0.0002	(0.0028)	−0.0090	(0.0026)
$\hat{\sigma}_1$	0.0215	n.c.	0.0291	n.c.
Implied ρ	−0.01	n.c.	−0.31	n.c.
	$N = 2258$		$N = 1036$	

For alcoholic beverages the inclusion of λ does not affect the results very much and

we obtain estimates that are pretty close to those reported in Table 7.8. The t-statistic on the coefficient for λ does not allow us to reject the null hypothesis of no correlation, while the estimation results imply an estimated correlation coefficient of only −0.01.

$$\rho = \frac{\hat{\sigma}_1}{\lambda}$$

For tobacco, on the other hand, we do find a significant impact of the sample selection term λ, with an implied estimated correlation coefficient of −0.31.

Panel GMM

$$y_{it} = x_{it}'\beta + u_{it}, \quad i = 1, \ldots, n,\ t = 1, \ldots, T$$

x_{it} contains endogenous variables and lagged dependent variables.
IV: $Z_i (T \times L)$

population moment:

$$E[Z_i' u_i] = 0$$

sample moment:

$$g_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} Z_i'(y_i - X_i\beta)$$

GMM estimator:

$$\hat{\beta}_{\text{PGMM}}(W) = (S_{xz}' W S_{xz})^{-1} S_{xz}' W s_{zy}$$

asymptotic normality of $\hat{\beta}_{\text{PGMM}}$:

$$\sqrt{n}\left(\hat{\beta}_{\text{PGMM}} - \beta\right) \xrightarrow{d} N(0,\ \text{Avar}(\hat{\beta}_{\text{PGMM}}))$$

$$\text{Avar}(\hat{\beta}_{\text{PGMM}}) = \left(\Sigma_{xz}' W \Sigma_{xz}\right)^{-1} \Sigma_{xz}' W S W \Sigma_{xz} \left(\Sigma_{xz}' W \Sigma_{xz}\right)^{-1}$$

$$\widehat{\text{Avar}}(\hat{\beta}_{\text{PGMM}}) = (S_{xz}' W S_{xz})^{-1} S_{xz}' \hat{W} \hat{S} \hat{W} S_{xz} (S_{xz}' W S_{xz})^{-1}$$

Panel-robust standard errors allowing for both heteroskedasticity and auto-correlation.

Set $W = S_{zz}^{-1}$, **One-Step GMM** or panel-2SLS estimator:

$$\hat{\beta}_{\text{2SLS}} = \left[X' Z (Z'Z)^{-1} Z' X\right]^{-1} X' Z (Z'Z)^{-1} Z' y = (\hat{X}'\hat{X})^{-1} \hat{X}' y$$

$$\widehat{\text{Avar}}(\hat{\beta}_{\text{2SLS}}) = \hat{\sigma}^2 (\hat{X}'\hat{X})^{-1}, \quad \hat{\sigma}^2 = \frac{1}{nT} \sum_{i,t} \hat{u}_{it}^2$$

Set $W = \hat{S}^{-1}$, $\hat{S} \to S$, **Two-step GMM** estimator (efficient):

$$\hat{\beta}_{\text{2SGMM}} = \left[X' Z \hat{S}^{-1} Z' X\right]^{-1} X' Z \hat{S}^{-1} Z' y$$

$$\widehat{\text{Avar}}(\hat{\beta}_{\text{2SGMM}}) = \left[X' Z \hat{S}^{-1} Z' X\right]^{-1}$$

Tests of Over-identifying Restrictions (OIR, Hansen's J):

$$OIR = J = n \cdot g(\hat{\beta}_{\text{2SGMM}})' \hat{S}^{-1} g(\hat{\beta}_{\text{2SGMM}}) \xrightarrow{d} \chi^2(L - K)$$

IV Selection

Exogeneity Assumption	Moment Condition	Instrument Vector[a]
Summation	$\mathrm{E}\left[\sum_t \mathbf{z}_{it} u_{it}\right] = \mathbf{0}$	$[\mathbf{z}_{it}]$
Contemporaneous	$\mathrm{E}\left[\mathbf{z}_{it} u_{it}\right] = \mathbf{0}$, all t	$[\mathbf{0}'_{r_1} \cdots \mathbf{0}'_{r_{t-1}} \ \mathbf{z}'_{it} \ \mathbf{0}'_{r_{t+1}} \cdots \mathbf{0}'_{r_T}]$
Weak	$\mathrm{E}\left[\mathbf{z}_{is} u_{it}\right] = \mathbf{0}$, $s \leq t$, all t	$[\mathbf{0}'_{r_1} \cdots \mathbf{0}'_{r_{t-1}} \ (\mathbf{z}^t_{it})' \ \mathbf{0}'_{r_{t+1}} \cdots \mathbf{0}'_{r_T}]$
Strong	$\mathrm{E}\left[\mathbf{z}_{is} u_{it}\right] = \mathbf{0}$, all s and t	$[\mathbf{0}'_{r_1} \cdots \mathbf{0}'_{r_{t-1}} \ (\mathbf{z}^T_{it})' \ \mathbf{0}'_{r_{t+1}} \cdots \mathbf{0}'_{r_T}]$

[a] The instrument vector is the tth row of $\mathbf{Z}_i$ in (22.11); $(\mathbf{z}^t_{it})' = [\mathbf{z}'_{i1} \ldots \mathbf{z}'_{it}]$, $(\mathbf{z}^T_{it})' = [\mathbf{z}'_{i1} \ldots \mathbf{z}'_{iT}]$; and $r_s = \dim[\mathbf{z}'_{is}]$ or $\dim[\mathbf{z}^s_{is}]$ or $\dim[\mathbf{z}^T_{is}]$.

Summation Assumption

$$\mathbf{Z}_i = \begin{bmatrix} \mathbf{z}'_{i1} \\ \mathbf{z}'_{i2} \\ \vdots \\ \mathbf{z}'_{i\mathrm{T}} \end{bmatrix}, \quad \mathbf{u}_i = \begin{bmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{iT} \end{bmatrix}$$

$$\mathrm{E}\left[\sum_{t=1}^{T} \mathbf{z}_{it} u_{it}\right] = \mathbf{0}$$

The sum of the expectations between the instruments and the error terms over the entire sample period is zero.

Contemporaneous Exogeneity Assumption(Stronger)

$$\mathbf{Z}_i = \begin{bmatrix} \mathbf{z}'_{i1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{z}'_{i2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{z}'_{iT} \end{bmatrix}, \quad \mathbf{u}_i = \begin{bmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{iT} \end{bmatrix}$$

$$E[\mathbf{z}_{it} u_{it}] = \mathbf{0}, \quad t = 1, \cdots, T$$

The instrument in each period is uncorrelated with the error term in the same period. If x includes time dummies then there are only $TK - (T-1)$ moment conditions available.

- How to construct IV: For period t, we use only the current-period instrument and set instruments from other periods to zero.

Weak Exogeneity Assumption

$$E[z_{is} u_{it}] = 0, \quad s \leq t$$

The current and lagged values of the instrument are uncorrelated with the current error term.

- How to construct IV: For period t, we use all instruments up to and including period t.

Strong Exogeneity Assumption

$$E[z_{is}u_{it}] = 0, \quad s,t = 1,\dots,T$$

All values of the instrument (past, present, and future) are uncorrelated with the error term in any period.

- How to construct IV: For any period t, we can use instruments from all T periods.

Redundant Instruments

Time-invariant instruments can be used only once.
Instruments that are the product of time dummies and a time-invariant regressor should be included only once.

Chamberlain's Optimal Distance Estimator

$$y_{it} = \alpha_i + x_{it}'\beta + u_{it}$$

Assume $E[x_{is}u_{it}] = 0$, for all s,t.

> To eliminate α_i, we typically apply **within transformation** or **first-differencing**, obtaining a transformed model ($\tilde{y}_{it} = \tilde{x}_{it}'\beta + \tilde{u}_{it}$).
>
> However, this transformation **loses information**, and if x_{it} is strongly exogenous, then pre-transformation values of $x_{is}(s \neq t)$ from other periods can also serve as valid instruments.
>
> Chamberlain's method aims to **more fully utilize these cross-period moment conditions**, thereby obtaining a more efficient estimator than the standard fixed effects estimator.

Chamberlain's model setting:

$$y_i = e\alpha_i + (I_T \otimes \beta')x_i + u_i$$

$e = [1,1,\dots,1]'$

Chamberlain assumes the fixed effect α_i can be linearly predicted by the explanatory variables x_i (values from all periods):

$$E^*[\alpha_i|x_i] = \mu + \sum_{t=1}^{T}\lambda_t'x_{it} = \mu + \lambda'x_i$$

Assume $E^*[u_i|\alpha_i, x_i] = 0$

$$
\begin{aligned}
E^*[y_i|x_i] &= eE^*[\alpha_i|x_i] + (I_T \otimes \beta')x_i \\
&= e\mu + (I_T \otimes \beta' + e\lambda')x_i \\
&= \Pi_0 + \Pi_1 x_i
\end{aligned}
$$

Step 1: Estimate the Reduced-Form Parameters

- run a multivariate OLS regression to get the reduced-form parameter $\hat{\Pi}$.
- estimate the variance-covariance matrix of $\hat{\Pi}$: $\hat{V}[\text{Vec}(\hat{\Pi})]$ (Vec for Vectorization).

Step 2: Minimum Distance Estimation

$$
J(\beta, \lambda) = \Big(\text{Vec}(\hat{\Pi} - I_T \otimes \beta' - e\lambda')\Big)' W \Big(\text{Vec}(\hat{\Pi} - I_T \otimes \beta' - e\lambda')\Big)
$$

no homoskedasticity assumption is needed.

Example

Table 21.2. *Hours and Wages: Standard Linear Panel Model Estimators*[a]

	POLS	Between	Within	First Diff	RE–GLS	RE–MLE
α	7.442	7.483	7.220	.001	7.346	7.346
β	.083	.067	.168	.109	.119	.120
Robust se[b]	(.030)	(.024)	(.085)	(.084)	(.051)	(.052)
Boot se	[.030]	[.019]	[.084]	[.083]	[.056]	[.058]
Default se	{.009}	{.020}	{.019}	{.021}	{.014}	{.014}
R^2	.015	.021	.016	.008	.014	.014
RMSE	.283	.177	.233	.296	.233	.233
RSS	427.225	0.363	259.398	417.944	288.860	288.612
TSS	433.831	17.015	263.677	420.223	293.023	292.773
σ_α	.000		.181		.161	.162
σ_ε	.283		.232		.233	.233
λ	0.000	–	1.000	–	.585	.586
N	5320	532	5320	4788	5320	5320

[a] Shown are pooled OLS (POLS), between, within, first-differences, random effects (RE) GLS and MLE linear panel regression of lnhrs on lnwg. Standard errors for the slope coefficients are panel robust in parentheses, panel bootstrap in square brackets, and default estimates that assume iid errors in curly braces. The R^2, root mean square error (RMSE), residual sum of squares (RSS), total sum of squares (TSS), and sample size come from the appropriate regression given in Section 21.2. The parameter λ is defined after (21.11).

[b] se, standard error.

		Base Case		Stacked	
	OLS	**2SLS**	**2SGMM**	**2SLS**	**2SGMM**
β_1	0.112	0.209	0.547	0.543	0.330
Panel se	(.096)	(.374)	(.327)	(.209)	(.110)
Het se	[.079]	[.423]	[–]	[.226]	[–]
Default se	{.023}	{.389}	{ – }	{.169}	{ – }
RMSE	.283	.296	.307	.307	.298
Instruments	5	9	9	72	72
OIR Test	–	–	5.45	–	69.51
dof	–	–	4	–	67
p-value	–	–	.244	–	.393
N	4256	4256	4256	4256	4256

[a] Differenced regression uses annual data from 1981–1988 for 532 men. Reported are β_1, the coefficient of Δ lnwg, and three estimated standard errors: panel robust in parentheses, heteroskedastic robust in square brackets, and usual default estimates that assume iid errors in curly braces. All regressions additionally include Δkids, Δage, Δagesq, and Δdisab as regressors but their coefficient estimates are not reported. The instruments are lnwg lagged twice and kids, age, agesq, and disab lagged both once and twice. For the base case there are 9 instruments and for stacked instruments there are $8 \times 9 = 72$ instruments. RMSE is the root mean square error of the residual. OIR is the over identifying restictions test statistic, dof is the degrees of freedom, and p-value is the p-value for this test.

$$\Delta \ln \text{hrs}_{it} = \beta_1 \Delta \ln \text{wg}_{it} + \beta_2 \Delta \text{kids}_{it} + \beta_3 \Delta \text{age}_{it} + \beta_4 \Delta \text{agesq}_{it} + \beta_5 \Delta \text{disab}_{it} + \Delta u_{it}$$

2SLS with Base-Case Instruments: 9 IV
$lnwg_{i,t-2}, kids_{i,t-1}, age_{i,t-1}, agesq_{i,t-1}, disab_{i,t-1}, kids_{i,t-2}, age_{i,t-2}, agesq_{i,t-2}, disab_{i,t-2}$

2SLS with Stacked Instruments: 72(=8×9) IV

The **two-step GMM estimator** is more efficient than 2SLS.

Test of Overidentifying Restrictions: see OIR Test and its p-value
H_0: All overidentifying moment conditions are valid. In other words: all instruments are exogenous.

Test of Weak Instruments:
H_0: The correlation between the instruments and the endogenous explanatory variable is weak./The coefficients of the instruments in the first-stage regression are jointly zero (or close to zero).

Chap1:

1. Assumptions

> For Ass 2, distinguish strict and weak exogeneity.
> If we assume that x is unstochastic(fixed regressor), then we don't need the exogeneity assumption.

2. proposition 1.1, 1.2

> don't need to proof but remember the conclusions!

3. hypothesis test

- T
- F(Wald test)

4. GLS

Chap2:

1. convergence in probability and distribution

> Lemmas can be used directly.
> Assumptions, WLLN, CLT need to be write down.

2. ergodic stationary, weak stationary, martingale difference stationary
3. OLS large sample properties

- consistency
- asymptotic normality
- asymptotic variance estimation

4. hypothesis test

- t: t distribution in finite sample, normal distribution in large sample
- F: F distribution in finite sample, Chi2 distribution in large sample

> heteroskedasticity is allowed.

Chap3:

1. definitions

> conditional and unconditional likelihood function
> normal distibution, binominal distibution, logistic distribution
> maximum likelihood estimation(only consider the linear model)

2. Cramer-Rao bound

> find the least variance's estimator in unbiased estimators.
> $\hat{\beta}$ can attain the bound, $\hat{\sigma}^2$ can't.

3. hypothesis test

- linear:
  - Wald:
    - W1: $\sim \chi^2$, if we know σ^2
    - W2: $\rightarrow \chi^2$, we use s^2 to subsitude the unknown σ^2
    - W3=W2/h: $\sim F$, since we use the unbiased $s^2 = SSR/(n-K)$
  - LR: connection with F test
- nonlinear:
  - $W \rightarrow \chi^2$
  - $F = W/h \rightarrow F$
  - $W_Z = \sqrt{W} \rightarrow N$

Chap4:

1. edogeneity from

- omitted variables
- measurement error
- simultaneous equations

> lead to unbiasedness and inconsistency

2. IV
   1. assumptions: correlation, exogeneity
   2. estimation, relations to OLS and GMM
   3. order condition and rank condition
   4. assumptions for large sample(don't need to proof)

5. 2SLS: when just indentified, 2SLS becomes IV.

Chap5:

1. moment/orthogonality condition: population moment, sample moment
2. GMM objective function
3. sample error
4. proposition 3.1, 3.2

- consistency
- asymptotic normality
- asymptotic variance estimation

5. efficient GMM: least variance($\hat{W} = \hat{S}^{-1}$)
6. hypothesis test(proposition 3.6, 3.7, 3.3)

> J test(over identifying restriction test) = GMM objective function
> H_0? df = K - L

7. homoskedasticity

- GMM to 2SLS
- J to Sargan

> asymptotic variance

Chap6:

1. Assumptions: versus SEGMM
2. MEGMM objective function, estimator $\tilde{\delta}$

> when use zipped matrix write down the dimensions.

3. special cases: FIVE, 3SLS, SUR

- Assumptions
- large sample properties(don't need to proof)
- relations

4. SUR versus OLS(Figure 4.1)

Chap9&10:

1. binary choice model
- logit
- probit

> definition, likelihood function, goodness of fit

2. multiple choice model
- orded(extended probit with mutiple hurdles)
- non-orded(extended logit, Gumble ditribution)
  IIA(odds ratio): idependence between two probabilities

3. count data: possion(estimate by MLE), equidispersion(NegBin I & II)

4. Tobit1

- cencered(3 equations, y=0 exists)

> logL1= part1 +part2

- truncated(2 eqution s)

5. Tobit2: 5 parts, sample selection bias=inverse Mill's ratio
6. Heckman 2-step estimation versus MLE

Chap11:

1. GMM conclusions: $\hat{\beta}_{PGMM}$, $\hat{V}[\hat{\beta}_{PGMM}]$
2. IV selection(4 exogeneity assumptions)
3. Chamberlain's optimal distance estimator