第 12 章　面 板 数 据

12.1　面板数据的特点

面板数据(panel data 或 longitudinal data)，指的是在一段时间内跟踪同一组个体(individual)的数据。

它既有横截面维度(n位个体)，又有时间维度(T个时期)。

一个$T=3$的面板数据结构如表 12.1。

表 12.1　面板数据的结构

	y	x_1	x_2	x_3
个体 1:　$t=1$				
个体 1:　$t=2$				
个体 1:　$t=3$				
个体 2:　$t=1$				
个体 2:　$t=2$				
个体 2:　$t=3$				
…….				
个体n:　$t=1$				
个体n:　$t=2$				
个体n:　$t=3$				

通常的面板数据T较小，而n较大，在使用大样本理论时让n趋于无穷大。这种面板数据称为**短面板**(short panel)。

反之，如果T较大，而n较小，则称为“长面板”(long panel)。在实践中，短面板较为常见。

在面板模型中，如果解释变量包含被解释变量的滞后值，则称为**动态面板**(dynamic panel)；反之，则称为**静态面板**(static panel)。

如果在面板数据中，每个时期在样本中的个体完全一样，则称为**平衡面板**(balanced panel)；

反之，则称为**非平衡面板**(unbalanced panel)。

面板数据的主要优点如下。

(1) 可以解决遗漏变量问题

遗漏变量偏差是普遍存在的问题。

遗漏变量常常是由于不可观测的个体差异或“异质性”(heterogeneity)造成的(比如个体能力)

如果这种个体差异“不随时间而改变”(time invariant)，则面板数据提供了解决遗漏变量问题的又一利器。

(2) 提供更多个体动态行为的信息

由于面板数据同时有横截面与时间两个维度，有时它可以解决单独的截面数据或时间序列所不能解决的问题。

比如，考虑如何区分规模效应与技术进步对企业生产效率的影响。对于截面数据，没有时间维度，故无法观测到技术进步。对于单个企业的时间序列数据来说，也无法区分其生产效率的提高究竟有多少是由于规模扩大，有多少是由于技术进步。

又比如，对于失业问题，截面数据能告诉我们在某个时点上哪些人失业，而时间序列能告诉我们某个人就业与失业的历史，但这两种数据均无法告诉我们是否失业的总是同一批人(低流转率)，还是失业的人群总在变动(高流转率)。

如果有面板数据，就可能解决上述问题。

(3) 样本容量较大

由于同时有截面维度与时间维度，通常面板数据的样本容量更大，从而可以提高估计的精确度。

面板数据也会带来一些问题。

样本数据通常不满足独立同分布的假定，因为同一个体在不同期的扰动项一般存在自相关。

面板数据的收集成本通常较高，不易获得。

12.2 面板数据的估计策略

估计面板数据的一个极端策略是将其看成是截面数据而进行**混合回归**(pooled regression)，即要求样本中每位个体都拥有完全相同的回归方程。

混合回归的缺点：忽略了个体不可观测的异质性(heterogeneity)，而该异质性可能与解释变量相关，导致估计不一致。

另一极端策略则是，为每位个体估计一个单独的回归方程。

分别回归的缺点：忽略了个体的共性，且可能没有足够大的样本容量(尤其对于短面板而言)。

实践中常采用折衷的估计策略，即假定个体的回归方程拥有相同的斜率，但可有不同截距项，以捕捉异质性(参见图 12.1)。

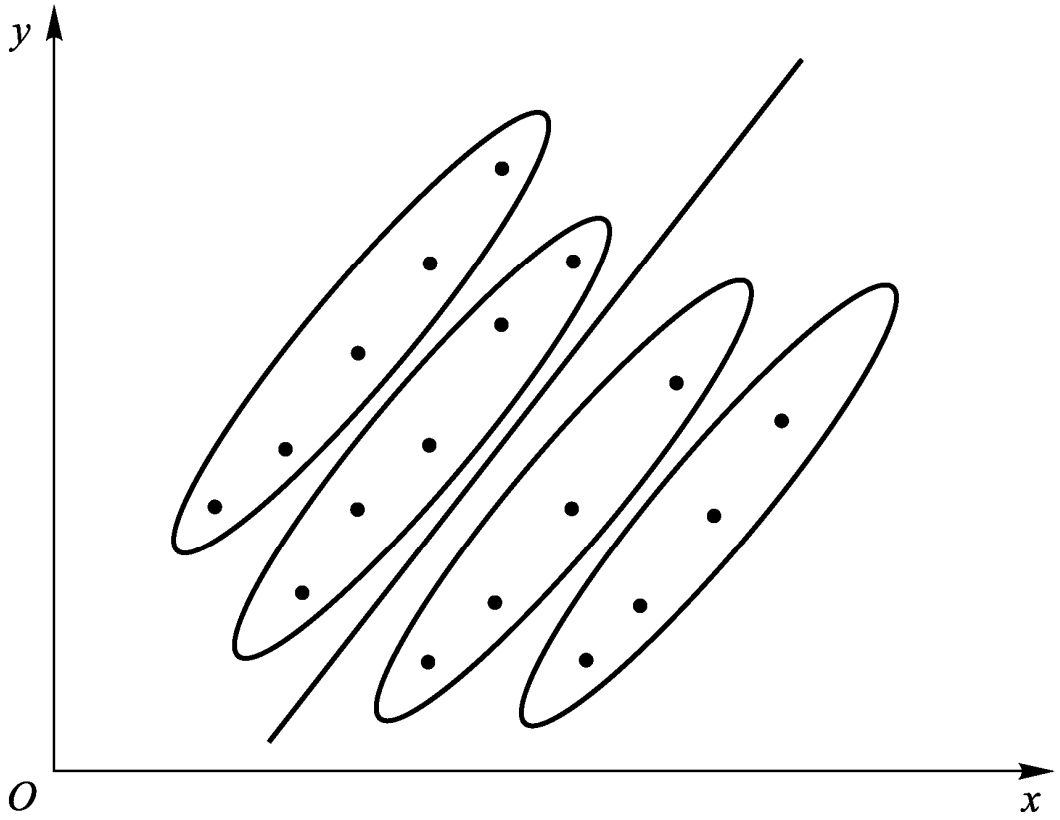

图 12.1　面板数据中不同个体的截距项可以不同

这种模型被称为**个体效应模型**(individual-specific effects model)：

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\delta} + u_i + \varepsilon_{it} \quad (i = 1, \cdots, n; t = 1, \cdots, T) \qquad (12.1)$$

$\mathbf{z}_i$为不随时间而变(time invariant)的个体特征(即$\mathbf{z}_{it} = \mathbf{z}_i, \forall t$)，比如性别。

$\mathbf{x}_{it}$可以随个体及时间而变(time-varying)。

扰动项由$(u_i + \varepsilon_{it})$两部分构成，称为“复合扰动项”(composite error term)。

不可观测的随机变量u_i是代表个体异质性的截距项，即**个体效应**(individual effects)。

在较早的文献中有时将u_i视为常数(待估参数)，但这也只是随机变量的特例，即退化的随机变量。

扰动项ε_{it}既随个体又随时间而变，称为“个殊性扰动项”(idiosyncratic error)。

一般假设$\{\varepsilon_{it}\}$为独立同分布，且与u_i不相关。

如果u_i与某个解释变量相关，则称之为**固定效应模型**(Fixed Effects Model，简记 FE)。此时，OLS 不一致。

如果u_i与所有解释变量$(\mathbf{x}_{it}, \mathbf{z}_i)$均不相关，则称之为**随机效应模型**(Random Effects Model，简记 RE)。

12.3 混合回归

如果所有个体都拥有完全一样的回归方程，则$u_1 = u_2 = \cdots = u_n$。将这些相同的个体效应统一记为α，则方程(12.1)可写为：

$$y_{it} = \alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\delta} + \varepsilon_{it} \qquad (12.2)$$

其中，$\mathbf{x}_{it}$不包括常数项。

可把所有数据放在一起，像对待横截面数据那样进行 OLS 回归，故称为**混合回归**(pooled regression)。

虽然通常可以假设不同个体之间的扰动项相互独立，但同一个体在不同时期的扰动项之间往往存在自相关。

每位个体不同时期的所有观测值即构成一个“聚类”(cluster)。

样本观测值可以分为不同的聚类，在同一聚类里的观测值互相相关，而不同聚类之间的观测值则不相关，称为**聚类样本**(cluster sample)。

对于聚类样本，仍可进行 OLS 估计，但需使用**聚类稳健的标准误**(cluster-robust standard errors)，在形式上也是一种夹心估计量，只是表达式更为复杂。

对于样本容量为nT的平衡面板，共有n个聚类，而每个聚类中包含T期观测值。

使用聚类稳健标准误的前提是，聚类中的观测值数目T较小，而聚类数目n较大($n \to \infty$)；此时，聚类稳健标准误是真实标准误的一致估计。

聚类稳健标准误更适用于时间维度T比截面维度n小的短面板。

由于在其推导过程中并未假定同方差，故聚类稳健的标准误也是异方差稳健的。

混合回归的基本假设是不存在个体效应，对此假设须进行检验。

由于个体效应以两种不同的形态存在(即固定效应与随机效应)，将分别介绍其检验方法。

12.4 固定效应模型：组内估计量

考虑以下固定效应模型：

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\delta} + u_i + \varepsilon_{it} \qquad (12.3)$$

其中，u_i与某解释变量相关，故 OLS 不一致。

解决方法是，通过模型变换，消掉个体效应u_i。在方程(12.3)中，给定个体 i，将方程两边对时间取平均可得

$$\bar{y}_i = \bar{\mathbf{x}}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\delta} + u_i + \bar{\varepsilon}_i \qquad (12.4)$$

其中，$\bar{y}_i \equiv \frac{1}{T}\sum_{t=1}^{T} y_{it}$，$\bar{\mathbf{x}}_i$与$\bar{\varepsilon}_i$的定义类似

将原方程(12.3)减去平均方程(12.4)，可得模型的离差形式：

$$y_{it}-\overline{y}_i=(\mathbf{x}_{it}-\overline{\mathbf{x}}_i)'\boldsymbol{\beta}+(\varepsilon_{it}-\overline{\varepsilon}_i) \qquad (12.5)$$

其中，$\mathbf{z}_i$与u_i都被消去。

定义$\tilde{y}_{it}\equiv y_{it}-\overline{y}_i$，$\tilde{\mathbf{x}}_{it}\equiv \mathbf{x}_{it}-\overline{\mathbf{x}}_i$，$\tilde{\varepsilon}_{it}\equiv \varepsilon_{it}-\overline{\varepsilon}_i$，则

$$\tilde{y}_{it}=\tilde{\mathbf{x}}_{it}'\boldsymbol{\beta}+\tilde{\varepsilon}_{it} \qquad (12.6)$$

只要新扰动项$\tilde{\varepsilon}_{it}$与新解释变量$\tilde{\mathbf{x}}_{it}$不相关，则可用 OLS 一致地估计$\boldsymbol{\beta}$，称为**固定效应估计量**(Fixed Effects Estimator)，记为$\hat{\boldsymbol{\beta}}_{\text{FE}}$。

由于$\hat{\boldsymbol{\beta}}_{\text{FE}}$主要使用了每位个体的组内离差信息，故也称为**组内估计量**(within estimator)。

即使个体特征u_i与解释变量$\mathbf{x}_{it}$相关，只要使用组内估计量，即可得到一致估计，这是面板数据的一大优势。

考虑到可能存在组内自相关，故应使用以每位个体为聚类的聚类稳健标准误。

在作离差变换的过程中，$\mathbf{z}_i'\boldsymbol{\delta}$也被消掉，故无法估计$\boldsymbol{\delta}$。

$\hat{\boldsymbol{\beta}}_{\text{FE}}$无法估计不随时间而变的变量之影响，这是 FE 的一大缺点。

为保证$(\varepsilon_{it}-\overline{\varepsilon}_i)$与$(\mathbf{x}_{it}-\overline{\mathbf{x}}_i)$不相关，须假定个体 i 满足严格外生性(比前定变量或同期外生的假定更强)，即$\mathrm{E}(\varepsilon_{it}\left|\mathbf{x}_{i1},\cdots,\mathbf{x}_{iT}\right.)=0$，因为$\overline{\mathbf{x}}_i$中包含了所有$(\mathbf{x}_{i1},\cdots,\mathbf{x}_{iT})$的信息。

12.5 固定效应模型：LSDV 法

对于方程(12.3)中的个体固定效应u_i，传统上将其视为个体 i 的待估参数；具体来说，可视u_i为个体 i 的截距项。

对于n位个体的n个不同截距项，可通过在方程(12.3)中引入$(n-1)$个个体虚拟变量来体现(如果没有截距项，则引入n个虚拟变量)，即估计以下模型：

$$y_{it} = \alpha + \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\delta} + \sum_{i=2}^{n} \gamma_i D_i + \varepsilon_{it} \qquad (12.7)$$

其中，个体虚拟变量D_2=1，如果为个体 2；否则，D_2= 0。其他个体虚拟变量$(D_3, \cdots, D_n)$的定义类似。

常数项α表示被遗漏虚拟变量D_1所对应的个体 1 的截距项，而个体 i $(i > 1)$的截距项则为$(\alpha + \gamma_i)$。

用 OLS 估计方程(12.7)，称为**最小二乘虚拟变量法**(Least Square Dummy Variable，简记 LSDV)。

LSDV 法的估计结果与上述组内估计量 FE 完全相同。

这正如线性回归与离差形式的回归在某种意义上是等价的(参见习题)。比如，

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad \Leftrightarrow \quad y_i - \overline{y} = \beta(x_i - \overline{x}) + (\varepsilon_i - \overline{\varepsilon}) \qquad (12.8)$$

如果做完 LSDV 后发现某些个体的虚拟变量不显著而删去，则 LSDV 的结果就不会与 FE 相同。

使用 LSDV 的好处是可以得到对个体异质性u_i的估计。

LSDV 法的缺点是，如果n很大，则须在回归方程中引入很多虚拟变量，可能超出 Stata 所允许的变量个数。

12.6 固定效应模型：一阶差分法

对于固定效应模型，还可对原方程(12.3)两边进行一阶差分，以消去个体效应u_i(但同时也把$\mathbf{z}_i'\boldsymbol{\delta}$消掉了)：

$$y_{it} - y_{i,t-1} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta} + (\varepsilon_{it} - \varepsilon_{i,t-1}) \qquad (12.9)$$

对上式使用 OLS 即得到**一阶差分估计量**(First Differencing Estimator)，记为$\hat{\boldsymbol{\beta}}_{\mathrm{FD}}$。

只要扰动项的一阶差分$(\varepsilon_{it} - \varepsilon_{i,t-1})$与解释变量的一阶差分$(\mathbf{x}_{it} - \mathbf{x}_{i,t-1})$不相关，则$\hat{\boldsymbol{\beta}}_{\mathrm{FD}}$一致。

此一致性条件比保证$\hat{\boldsymbol{\beta}}_{\text{FE}}$一致的严格外生性假定更弱，这是$\hat{\boldsymbol{\beta}}_{\text{FD}}$的主要优点。

可以证明(参见习题)，如果$T=2$，则$\hat{\boldsymbol{\beta}}_{\text{FD}}=\hat{\boldsymbol{\beta}}_{\text{FE}}$。

但对于$T>2$，如果$\{\varepsilon_{it}\}$为独立同分布的，则组内估计量$\hat{\boldsymbol{\beta}}_{\text{FE}}$比一阶差分估计量$\hat{\boldsymbol{\beta}}_{\text{FD}}$更有效率。

在实践上，主要使用$\hat{\boldsymbol{\beta}}_{\text{FE}}$，较少用$\hat{\boldsymbol{\beta}}_{\text{FD}}$。

12.7 时间固定效应

个体固定效应模型解决了不随时间而变(time invariant)但随个体而异的遗漏变量问题。

但还可能存在不随个体而变(individual invariant)，但随时间而变(time varying)的遗漏变量问题；比如，企业经营的宏观经济环境。

在个体固定效应模型(12.3)中加入时间固定效应(λ_t)：

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\delta} + \lambda_t + u_i + \varepsilon_{it} \qquad (12.10)$$

其中，λ_t随时间而变，但不随个体而变；故只有下标 t，而没有下标 i。

可视λ_t为第t期特有的截距项，并解释为“第t期”对被解释变量y的效应；故称$\{\lambda_1, \cdots, \lambda_T\}$为**时间固定效应**(time fixed effects)。

可使用 LSDV 法来估计，即对每个时期定义一个虚拟变量，然后把$(T-1)$个时间虚拟变量包括在回归方程中(未包括的时间虚拟变量即为基期)，比如

$$y_{it}=\alpha+\mathbf{x}_{it}'\boldsymbol{\beta}+\mathbf{z}_i'\boldsymbol{\delta}+\sum_{t=2}^{T}\gamma_t D_t+u_i+\varepsilon_{it} \qquad (12.11)$$

其中，时间虚拟变量$D_2=1$，如果$t=2$；否则，$D_2=0$。其他时间虚拟变量$(D_3,\cdots,D_T)$的定义类似。

常数项α表示被遗漏虚拟变量D_1所对应的第 1 期截距项，而第t期 $(t>1)$的截距项则为$(\alpha+\gamma_t)$。

由于方程(12.11)既考虑了个体固定效应，又考虑了时间固定效应，故称为**双向固定效应**(Two-way FE)。

如果仅考虑个体固定效应，则称为**单向固定效应**(One-way FE)。

有时为节省参数(比如，时间维度T较大)，可引入一个时间趋势项，以替代上述$(T-1)$个时间虚拟变量：

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\delta} + \gamma t + u_i + \varepsilon_{it} \qquad (12.12)$$

上式隐含的假定是，每个时期的时间效应相等，每期均增加γ。

如果此假定不太可能成立，则应在方程中加入时间虚拟变量。

可通过检验这些时间虚拟变量的联合显著性来判断是否应使用双向固定效应模型。

12.8 随机效应模型

考虑以下随机效应模型：

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\delta} + u_i + \varepsilon_{it} \qquad (12.13)$$

其中，个体效应u_i与解释变量$\{\mathbf{x}_{it}, \mathbf{z}_i\}$均不相关，故 OLS 一致。

由于扰动项由$(u_i + \varepsilon_{it})$组成，不是球型扰动项，故 OLS 不是最有效率的。

假设不同个体之间的扰动项互不相关。

即便如此，由于u_i的存在，同一个体不同时期的扰动项之间仍存在自相关。

对于$t \neq s$，可以证明

$$\begin{aligned}\mathrm{Cov}(u_i+\varepsilon_{it}, u_i+\varepsilon_{is}) &= \mathrm{Cov}(u_i, u_i) + \underbrace{\mathrm{Cov}(u_i, \varepsilon_{is})}_{=0} + \underbrace{\mathrm{Cov}(\varepsilon_{it}, u_i)}_{=0} + \underbrace{\mathrm{Cov}(\varepsilon_{it}, \varepsilon_{is})}_{0} \\ &= \mathrm{Var}(u_i) \equiv \sigma_u^2 \neq 0 \end{aligned} \quad (12.14)$$

其中，$\sigma_u^2 \equiv \mathrm{Var}(u_i)$为个体效应$u_i$的方差(不随$i$变化)。

在上式中，如果$t = s$，则

$$\mathrm{Var}(u_i + \varepsilon_{it}) = \sigma_u^2 + \sigma_\varepsilon^2 \qquad (12.15)$$

其中，$\sigma_\varepsilon^2 \equiv \mathrm{Var}(\varepsilon_{it})$为$\varepsilon_{it}$的方差(不随$i, t$变化)。

当$t \neq s$时，个体i扰动项的自相关系数为

$$\rho \equiv \mathrm{Corr}(u_i + \varepsilon_{it},\ u_i + \varepsilon_{is}) \equiv \frac{\mathrm{Cov}(u_i + \varepsilon_{it}, u_i + \varepsilon_{is})}{\mathrm{Var}(u_i + \varepsilon_{it})} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2} \qquad (12.16)$$

自相关系数ρ越大，则复合扰动项$(u_i + \varepsilon_{it})$中个体效应的部分(u_i)越重要。

Stata 记ρ为“rho”。

由于方程(12.13)的扰动项$(u_i + \varepsilon_{it})$存在组内自相关，故 OLS 不是最有效率的。

可使用广义最小二乘法(GLS)对原模型进行转换，使得变换后的扰动项不再有自相关。

首先定义

$$\theta \equiv 1 - \frac{\sigma_\varepsilon}{(T\sigma_u^2 + \sigma_\varepsilon^2)^{1/2}} \qquad (12.17)$$

其中，T 为面板数据的时间维度。

显然，$0 \le \theta \le 1$。

给定个体i，将方程(12.13)两边对时间进行平均，同乘以θ可得

$$\theta\overline{y}_i = \theta\overline{\mathbf{x}}_i'\boldsymbol{\beta} + \theta\mathbf{z}_i'\boldsymbol{\delta} + \theta u_i + \theta\overline{\varepsilon}_i \qquad (12.18)$$

将原方程(12.13)减去方程(12.18)可得**准离差**(quasi-demeaned)模型：

$$y_{it} - \theta\overline{y}_i = (\mathbf{x}_{it} - \theta\overline{\mathbf{x}}_i)'\boldsymbol{\beta} + (1-\theta)\mathbf{z}_i'\boldsymbol{\delta} + \underbrace{\left[(1-\theta)u_i + (\varepsilon_{it} - \theta\overline{\varepsilon}_i)\right]}_{\text{扰动项}}$$

(12.19)

由于$0 \le \theta \le 1$，故$(y_{it} - \theta\overline{y}_i)$只是减去平均值$\overline{y}_i$的一部分，故名“广义离差”。

可以证明，广义离差方程(12.19)的扰动项$\left[(1-\theta)u_i+(\varepsilon_{it}-\theta\overline{\varepsilon}_i)\right]$不再有自相关(尽管它仍包含$u_i$)，对此方程进行OLS估计即为GLS估计量。

但θ通常未知(取决于u_i与ε_{it}的方差)，故须先估计$\hat{\theta}$，再进行FGLS估计。可用下式来估计$\hat{\theta}$：

$$\hat{\theta}\equiv 1-\frac{\hat{\sigma}_\varepsilon}{(T\hat{\sigma}_u^2+\hat{\sigma}_\varepsilon^2)^{1/2}} \qquad (12.20)$$

其中，$\hat{\sigma}_u$与$\hat{\sigma}_\varepsilon$分别为σ_u与σ_ε的样本估计值。

Stata分别记$\hat{\sigma}_u$、$\hat{\sigma}_\varepsilon$与$\hat{\theta}$为“`sigma_u`”、“`sigma_e`”与“`theta`”。

对于随机效应模型，由于 OLS 是一致的，且其扰动项为$(u_i+\varepsilon_{it})$，故可用 OLS 的残差来估计$(\sigma_u^2+\sigma_\varepsilon^2)$。

FE 也是一致的，且其扰动项为$(\varepsilon_{it}-\overline{\varepsilon}_i)$，故可用 FE 的残差来估计$\sigma_\varepsilon^2$。

由此得到$\hat{\theta}$，再使用 FGLS 估计原模型，即可得到**随机效应估计量**(Random Effects Estimator)，记为$\hat{\boldsymbol{\beta}}_{\mathrm{RE}}$。

对于随机效应模型，如果假设扰动项服从正态分布，则可写出样本的似然函数，然后进行最大似然估计(MLE)。

12.9 组间估计量

对于随机效应模型，还可以使用“组间估计量”。

如果每位个体的时间序列数据较不准确或噪音较大，可对每位个体取时间平均值，然后用平均值来作横截面回归：

$$\overline{y}_i = \overline{\mathbf{x}}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\delta} + u_i + \overline{\varepsilon}_i \quad (i = 1, \cdots, n) \qquad (12.21)$$

对上式使用 OLS，即为**组间估计量**(Between Estimator)，记为 $\hat{\boldsymbol{\beta}}_{\mathrm{BE}}$。

由于$\{\bar{\mathbf{x}}_i, \mathbf{z}_i\}$中包含了$\{\mathbf{x}_{it}, \mathbf{z}_i\}$的信息，如果$u_i$与解释变量$\{\mathbf{x}_{it}, \mathbf{z}_i\}$相关，则$\hat{\boldsymbol{\beta}}_{\text{BE}}$不一致。

不能在固定效应模型下使用组间估计法。

即使在随机效应模型下，由于面板数据被压缩为截面数据，损失了较多信息量，故组间估计法也不常用。

12.10 拟合优度的度量

对于面板模型，如果使用混合回归，则可直接用混合回归的R^2衡量拟合优度。

如果使用固定效应或随机效应，拟合优度的度量略为复杂。

对于有常数项的线性回归模型，其拟合优度R^2等于被解释变量y与预测值$\hat{y}$之间相关系数的平方，即$R^2=[\text{corr}(y,\hat{y})]^2$。

给定估计量$(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\delta}})$，Stata 提供了以下三种$R^2$。

(1) 对应于原模型(12.1)，称$[\text{Corr}(y_{it},\mathbf{x}_{it}'\hat{\boldsymbol{\beta}}+\mathbf{z}_i'\hat{\boldsymbol{\delta}})]^2$为**整体**$R^2$($R^2$overall)，衡量估计量$(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\delta}})$对原模型的拟合优度。

(2) 对应于组内模型(12.6)，称$[\text{Corr}(\tilde{y}_{it},\tilde{\mathbf{x}}_{it}'\hat{\boldsymbol{\beta}})]^2$为**组内**$R^2$($R^2$within)，衡量估计量$(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\delta}})$对组内模型的拟合优度。

(3) 对应于组间模型(12.21)，称$[\text{Corr}(\overline{y}_i,\overline{\mathbf{x}}_i'\hat{\boldsymbol{\beta}}+\mathbf{z}_i'\hat{\boldsymbol{\delta}})]^2$为**组间**$R^2$($R^2$between)，衡量估计量$(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\delta}})$对组间模型的拟合优度。

无论固定效应、随机效应还是组间回归，都可计算以上三种R^2。

对于固定效应模型，建议使用组内R^2。

对于组间回归模型，建议使用组间R^2。

对于随机效应模型，这三种R^2都只是相应的相关系数平方而已(并非随机效应模型的R^2)。

12.11 非平衡面板

在面板数据中，如果每个时期在样本中的个体完全一样，则称为“平衡面板数据” (balanced panel)。

但有时某些个体的数据可能缺失(比如，个体死亡、企业倒闭或被兼并、个体不再参与调查)，或者新个体在后来才加入到调查中来。

如果每个时期观测到的个体不完全相同，称为“非平衡面板”(unbalanced panel)或“不完全面板”(incomplete panel)。

非平衡面板数据并不影响计算离差形式的组内估计量(within estimator)，故固定效应模型的估计可照样进行。

对于随机效应模型而言，非平衡面板数据也没有实质性影响。

假设个体i的时间维度为T_i，则只要在做广义离差变换时，为每位个体定义

$$\hat{\theta}_i \equiv 1 - \frac{\hat{\sigma}_\varepsilon}{(T_i\hat{\sigma}_u^2 + \hat{\sigma}_\varepsilon^2)^{1/2}} \qquad (12.22)$$

即可照常进行 FGLS 估计。

非平衡面板数据使得估计量及其协方差矩阵的数学表达式更加复杂，但这些都由 Stata 在幕后进行。

非平衡面板可能出现的最大问题是，那些原来在样本中但后来丢掉的个体，如果其“丢掉”的原因是内生的(即与扰动项相关)，则会导致样本不具有代表性(不再是随机样本)，从而导致估计量不一致。

比如，低收入的人群更容易从面板数据中丢掉。

如果从非平衡面板数据中提取一个平衡的面板数据子集，然后进行数据处理，则必然会损失样本容量，降低估计效率。

如果人为“丢掉”的个体并非完全随机，则同样会破坏样本的随机性。

12.12 究竟该用固定效应还是随机效应模型

在处理面板数据时，究竟应该使用固定效应还是随机效应模型是一个根本问题。

希望检验原假设“$H_0:u_i$与$\mathbf{x}_{it},\mathbf{z}_i$不相关”(即随机效应模型为正确模型)。

无论原假设成立与否，FE 都是一致的。

如果原假设成立，则 RE 一致且比 FE 更有效率。

如果原假设不成立，则 RE 不一致。

如果H_0成立，则 FE 与 RE 估计量将共同收敛于真实的参数值，二者的差距将在大样本下消失，故$(\hat{\boldsymbol{\beta}}_{\text{FE}} - \hat{\boldsymbol{\beta}}_{\text{RE}}) \xrightarrow{p} \mathbf{0}$。

反之，如果二者的差距过大，则倾向于拒绝原假设。

以二次型度量此距离，**豪斯曼检验**(Hausman, 1978)的统计量为

$$(\hat{\boldsymbol{\beta}}_{\text{FE}} - \hat{\boldsymbol{\beta}}_{\text{RE}})' \left[\widehat{\text{Var}(\hat{\boldsymbol{\beta}}_{\text{FE}} - \hat{\boldsymbol{\beta}}_{\text{RE}})} \right]^{-1} (\hat{\boldsymbol{\beta}}_{\text{FE}} - \hat{\boldsymbol{\beta}}_{\text{RE}}) \xrightarrow{d} \chi^2(K) \quad (12.23)$$

其中，K 为$\hat{\boldsymbol{\beta}}_{\mathrm{FE}}$的维度，即$\mathbf{x}_{it}$中所包含的随时间而变的解释变量个数(因为$\hat{\boldsymbol{\beta}}_{\mathrm{FE}}$无法估计不随时间而变的解释变量系数)。

如果该统计量大于临界值，则拒绝H_0。

此检验的缺点是，为了计算$\widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{FE}}-\hat{\boldsymbol{\beta}}_{\mathrm{RE}})}$，它假设在$H_0$成立的情况下，$\hat{\boldsymbol{\beta}}_{\mathrm{RE}}$是最有效率的(fully efficient)。

如果扰动项存在异方差，则$\hat{\boldsymbol{\beta}}_{\mathrm{RE}}$并非最有效率的估计量。

因此，传统的豪斯曼检验并不适用于异方差的情形，须使用异方差稳健的豪斯曼检验。

12.13　面板模型的 Stata 命令及实例

1．面板数据的设定

设定面板数据的 Stata 命令为

```
. xtset panelvar timevar
```

命令“xtset”告诉 Stata 你的数据为面板数据，其中面板(个体)变量“panelvar”的取值必须为整数且不重复，相当于将样本中每位个体进行编号；而“timevar”为时间变量。

假如“panelvar”本来是字符串(比如，国家名字 country)，则可使用以下命令将其转换为数字型变量：

```
. encode country, gen(cntry)
```

选择项“gen(cntry)”表示将新生成的数字型变量记为 cntry。这样，变量 cntry 就以“1, 2, 3, …”来指代不同的国家。

显示面板数据统计特性的 Stata 命令包括

```
. xtdes
```
(显示面板数据的结构，是否为平衡面板)

```
. xtsum
```
(显示组内、组间与整体的统计指标)

```
. xtline varname
```
(对每位个体分别显示该变量的时间序列图；如果希望将所有个体的时间序列图叠放在一起，可加上选择项 overlay)

以数据集 lin_1992.dta 为例，取自 Lin (1992)对家庭联产承包责任制(household responsibility system)与中国农业增长的经典研究。

该省际面板包含中国 28 个省 1970—1987 年有关种植业的数据。

被解释变量：“种植业产值对数”(*ltvfo*，1980 年不变价格)。

解释变量：耕地面积对数(*ltlan*，千亩)，种植业劳动力对数(*ltwlab*)，机械动力与畜力对数(*ltpow*，千马力)，化肥使用量对数(*ltfer*，千吨)，截止年底采用家庭联产承包制的生产队比重(*hrs*)，农村消费者价格与农村工业投入品价格之比的一阶滞后(mipric1，1950 年=100)，超额收购价格与农村工业投入品价格之比(*giprice*，1950 年=100)，复种指数(*mci*，播种面积除以耕地面积)，非粮食作物占播种面积比重(*ngca*)，时间趋势(*t*)，*province*(省)，*year*(年)。

为解决异方差问题，Lin (1992)将种植业产量、耕地面积、种植业劳动力、机械动力与畜力、化肥使用量这些传统的投入与产出变量都除以每省的生产队数目(*team*)。

两个价格变量 *mipric1* 与 *giprice* 为全国性指标，在各省都一样，只随时间变化。

首先，设定 *province* 与 *year* 为面板(个体)变量及时间变量：

```
. use lin_1992.dta,clear
. xtset province year
```

```
Panel variable: province (strongly balanced)
 Time variable: year, 70 to 87
         Delta: 1 unit
```

这是一个平衡的面板数据(strongly balanced)。

其次，显示数据集的结构：

```
. xtdes
```

```
province:  1, 2, ..., 28                                        n =         28
    year:  70, 71, ..., 87                                      T =         18
           Delta(year) = 1 unit
           Span(year)  = 18 periods
           (province*year uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%      max
                        18      18      18        18        18      18       18

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+------------------
       28    100.00  100.00 |  111111111111111111
 ---------------------------+------------------
       28    100.00         |  XXXXXXXXXXXXXXXXXX
```

$n=28$，而$T=18$。由于n大而T小，故这是一个短面板。

再次，显示数据集中以上变量的统计特征：

```
. xtsum ltvfo ltlan ltwlab ltpow ltfer hrs mipric1 giprice mci ngca
```

Variable		Mean	Std. dev.	Min	Max	Observations	
ltvfo	overall	7.647758	.5331999	5.51	9.33	N =	504
	between		.4611992	6.982222	8.977222	n =	28
	within		.2806888	5.61498	8.471647	T =	18
ltlan	overall	5.837877	.8084866	4.57	7.76	N =	504
	between		.8143036	4.617222	7.697778	n =	28
	within		.1138892	4.758988	6.163988	T =	18
ltwlab	overall	3.19752	.4193496	.98	3.86	N =	504
	between		.3195715	2.303889	3.646111	n =	28
	within		.2778123	1.618631	4.053631	T =	18
ltpow	overall	2.692778	.9463811	.2	5.04	N =	504
	between		.7702036	1.475	4.180556	n =	28
	within		.5678668	.31	3.909444	T =	18

ltfer	overall	2.15119	.7903761	-.23	3.98	N =	504
	between		.5624935	1.081111	3.649444	n =	28
	within		.564791	.4173016	3.510079	T =	18
hrs	overall	.3497479	.4526283	0	1	N =	476
	between		.0453814	.2123529	.4094118	n =	28
	within		.4504245	-.0596639	1.053866	T =	17
mipric1	overall	2.248889	.2431379	1.76	2.73	N =	504
	between		0	2.248889	2.248889	n =	28
	within		.2431379	1.76	2.73	T =	18
giprice	overall	2.858889	.4537578	2.39	3.56	N =	504
	between		0	2.858889	2.858889	n =	28
	within		.4537578	2.39	3.56	T =	18
mci	overall	1.538452	.4931854	.85	2.55	N =	504
	between		.4972044	.8666667	2.487222	n =	28
	within		.0661412	1.323452	1.880119	T =	18
ngca	overall	.199623	.076145	.06	.91	N =	504
	between		.0631671	.1144444	.3466667	n =	28
	within		.0440777	.1151786	.8951786	T =	18

除 *hrs* 外，所有变量的观测样本均为$28\times18=504$。

关键变量 *hrs* 的样本容量仅为$28\times17=476$，因为缺失 1980 年的 *hrs* 观测数据。

看一下被解释变量 *ltvfo* 在 28 个省的时间趋势图，结果如图 12.2。

```
. xtline ltvfo
```

图 12.2　28 省种植业产值的时间趋势图

虽然不同省的种植业产值均随时间而增长，但变化的趋势与时机不尽相同。

种植业产值的这些省际差异有助于估计决定种植业产值的因素。

2. 混合回归

作为参照系，首先进行混合回归。其 Stata 命令的基本格式为

```
. reg y x1 x2 x3,vce(cluster id)
```

其中，“id”指用来确定每位个体的变量，而选择项“vce(cluster id)”表示以变量 id 作为聚类变量来计算聚类稳健的标准误。

```
. reg ltvfo ltlan ltwlab ltpow ltfer hrs mipric1 giprice mci ngca,vce(cluster province)
```

其中，选择项“vce(cluster province)”表示，使用以“province”为聚类变量的聚类稳健标准误。

将此结果储存，并记为“OLS”。

```
. estimates store OLS
```

Linear regression

Number of obs	= 476
F(9, 27)	= 81.39
Prob > F	= 0.0000
R-squared	= 0.8685
Root MSE	= .19689

(Std. err. adjusted for 28 clusters in province)

ltvfo	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
ltlan	.693795	.115024	6.03	0.000	.4577853	.9298048
ltwlab	.2650224	.0566294	4.68	0.000	.1488285	.3812164
ltpow	-.0291884	.0670385	-0.44	0.667	-.1667401	.1083633
ltfer	.3110617	.0531318	5.85	0.000	.2020443	.4200792
hrs	.2286926	.0489458	4.67	0.000	.1282642	.329121
mipric1	.0122048	.0547799	0.22	0.825	-.1001943	.1246039
giprice	-.0538892	.0274468	-1.96	0.060	-.1102054	.002427
mci	.6949202	.1689692	4.11	0.000	.3482241	1.041616
ngca	.3053056	.5222639	0.58	0.564	-.7662914	1.376903
_cons	1.080587	.8269888	1.31	0.202	-.6162544	2.777427

关键变量 *hrs* 在 1%水平上显著为正。

如果使用普通标准误，则可输入命令：

```
. reg ltvfo ltlan ltwlab ltpow ltfer hrs mipric1 giprice mci ngca
```

Source	SS	df	MS	Number of obs	=	476
				F(9, 466)	=	342.09
Model	119.355964	9	13.2617737	Prob > F	=	0.0000
Residual	18.0652415	466	.038766613	R-squared	=	0.8685
				Adj R-squared	=	0.8660
Total	137.421205	475	.2893078	Root MSE	=	.19689

ltvfo	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
ltlan	.693795	.0368914	18.81	0.000	.6213008	.7662892
ltwlab	.2650224	.0238406	11.12	0.000	.2181741	.3118708
ltpow	-.0291884	.0331891	-0.88	0.380	-.0944073	.0360305
ltfer	.3110617	.0206459	15.07	0.000	.2704911	.3516324
hrs	.2286926	.0307121	7.45	0.000	.1683412	.289044
mipric1	.0122048	.0533863	0.23	0.819	-.0927028	.1171125
giprice	-.0538892	.0271999	-1.98	0.048	-.1073389	-.0004395
mci	.6949202	.0522416	13.30	0.000	.592262	.7975784
ngca	.3053056	.1952732	1.56	0.119	-.0784195	.6890306
_cons	1.080587	.2832576	3.81	0.000	.5239661	1.637207

对比聚类稳健标准误与普通标准误可知，前者均大于后者。

由于同一省不同年之间的扰动项一般存在自相关，而默认的普通标准误计算方法假设扰动项为独立同分布的，故普通标准误的估计并不准确。

3．固定效应

由于每个省的“省情”不同，可能存在不随时间而变的遗漏变量，故考虑使用固定效应模型(FE)。

固定效应模型(组内估计量)的 Stata 命令格式为

```
. xtreg y x1 x2 x3,fe r
```

选择项“fe”表示“fixed effects”(固定效应估计量)，默认为“re”表示“random effects”(随机效应估计量)。

选择项“r”表示使用聚类稳健标准误；如果使用选择项“vce(cluster id)”也能达到完全相同的效果。

LSDV 法的 Stata 命令为

```
. reg y x1 x2 x3 i.id,vce(cluster id)
```

“id”表示用来确定个体的变量

“i.id”则表示根据变量“id”而生成的虚拟变量。

选择项“vce(cluster id)”表示使用聚类稳健的标准误。

首先使用组内估计量，并记其估计结果为“FE_robust”：

```
. xtreg ltvfo ltlan ltwlab ltpow ltfer hrs mipric1 giprice mci ngca,fe r
. estimates store FE_robust
```

Fixed-effects (within) regression — Number of obs = 476
Group variable: province — Number of groups = 28

R-squared:
Within = 0.8746
Between = 0.6483
Overall = 0.6993

Obs per group:
min = 17
avg = 17.0
max = 17

F(9, 27) = 274.25
corr(u_i, Xb) = -0.3877 — Prob > F = 0.0000

(Std. err. adjusted for 28 clusters in province)

ltvfo	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
ltlan	.6370234	.1681335	3.79	0.001	.2920421	.9820048
ltwlab	.1387786	.0624585	2.22	0.035	.0106242	.2669329
ltpow	.0577152	.0755568	0.76	0.452	-.0973146	.2127451
ltfer	.1826281	.043592	4.19	0.000	.0931846	.2720716
hrs	.2134022	.0391104	5.46	0.000	.1331542	.2936501
mipric1	.0543577	.0590331	0.92	0.365	-.0667682	.1754837
giprice	-.0151451	.0245968	-0.62	0.543	-.0656135	.0353233
mci	.1943697	.0770515	2.52	0.018	.0362731	.3524663
ngca	.7562031	.3821261	1.98	0.058	-.0278549	1.540261
_cons	2.337895	.8552224	2.73	0.011	.583124	4.092667
sigma_u	.30549743					
sigma_e	.10589274					
rho	.89273901	(fraction of variance due to u_i)				

输出结果中包括常数项(_cons)，是所有个体效应u_i的平均值。

上表最后一行显示，“rho=0.89”，故复合扰动项$(u_i+\varepsilon_{it})$的方差主要来自个体效应u_i的变动。

究竟应该使用混合回归还是个体固定效应模型呢？

在使用命令“xtreg,fe”时，如果不加选择项“r”(将估计结果记为“FE”)，则输出结果还包含一个F检验，其原假设为“H_0：所有$u_i=0$”，即混合回归是可以接受的：

```
. xtreg ltvfo ltlan ltwlab ltpow ltfer hrs mipric1 giprice mci ngca,fe

. estimates store FE
```

```
Fixed-effects (within) regression               Number of obs     =       476
Group variable: province                        Number of groups  =        28

R-squared:                                      Obs per group:
     Within  = 0.8746                                         min =        17
     Between = 0.6483                                         avg =      17.0
     Overall = 0.6993                                         max =        17

                                                F(9, 439)         =    340.20
corr(u_i, Xb) = -0.3877                         Prob > F          =    0.0000
```

ltvfo	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
ltlan	.6370234	.0673191	9.46	0.000	.5047156	.7693312
ltwlab	.1387786	.0261554	5.31	0.000	.0873732	.190184
ltpow	.0577152	.0332508	1.74	0.083	-.0076352	.1230657
ltfer	.1826281	.0219921	8.30	0.000	.1394053	.225851
hrs	.2134022	.0223886	9.53	0.000	.1694	.2574043
mipric1	.0543577	.0421659	1.29	0.198	-.0285145	.1372299
giprice	-.0151451	.0187457	-0.81	0.420	-.0519876	.0216975
mci	.1943697	.0876884	2.22	0.027	.0220285	.366711
ngca	.7562031	.2168141	3.49	0.001	.3300804	1.182326
_cons	2.337895	.385253	6.07	0.000	1.580726	3.095065
sigma_u	.30549743					
sigma_e	.10589274					
rho	.89273901	(fraction of variance due to u_i)				

```
F test that all u_i=0: F(27, 439) = 43.41                     Prob > F = 0.0000
```

对于原假设“H_0：所有$u_i = 0$”，由于p值为 0.0000，故强烈拒绝原假设，即认为 FE 明显优于混合回归，应该允许每位个体拥有自己的截距项。

由于未使用聚类稳健标准误，故此F检验并不有效，因为普通标准误均小于聚类稳健标准误。

进一步通过 LSDV 法来考察(将估计结果记为“LSDV”)：

```
. reg ltvfo ltlan ltwlab ltpow ltfer hrs mipric1 giprice mci ngca i.province,vce(cluster province)

. estimates store LSDV
```

Linear regression

Number of obs	= 476
F(8, 27)	= .
Prob > F	= .
R-squared	= 0.9642
Root MSE	= .10589

(Std. err. adjusted for 28 clusters in province)

ltvfo	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
ltlan	.6370234	.1732267	3.68	0.001	.2815916	.9924553
ltwlab	.1387786	.0643506	2.16	0.040	.0067421	.2708151
ltpow	.0577152	.0778457	0.74	0.465	-.1020109	.2174414
ltfer	.1826281	.0449126	4.07	0.000	.0904751	.2747811
hrs	.2134022	.0402952	5.30	0.000	.1307233	.296081
mipric1	.0543577	.0608214	0.89	0.379	-.0704375	.1791529
giprice	-.0151451	.0253419	-0.60	0.555	-.0671423	.0368522
mci	.1943697	.0793856	2.45	0.021	.0314839	.3572555
ngca	.7562031	.3937018	1.92	0.065	-.0516063	1.564013
province						
beijing	-.1816498	.1247829	-1.46	0.157	-.4376832	.0743835
fujian	.051657	.0501916	1.03	0.313	-.0513277	.1546418
gansu	-.8165674	.1380808	-5.91	0.000	-1.099886	-.5332489

guangdong	-.010488	.055811	-0.19	0.852	-.1250028	.1040268
guangxi	-.2304637	.0570853	-4.04	0.000	-.3475932	-.1133342
guizhou	-.2350768	.0615353	-3.82	0.001	-.3613369	-.1088167
hebei	-.2923854	.0997217	-2.93	0.007	-.4969974	-.0877733
heilongjiang	-.1410195	.2892268	-0.49	0.630	-.7344638	.4524249
henan	-.0904581	.0435714	-2.08	0.048	-.1798593	-.0010569
hubei	.1118905	.0340584	3.29	0.003	.0420085	.1817725
hunan	-.0373775	.0607647	-0.62	0.544	-.1620563	.0873014
jiangsu	.1150954	.0342058	3.36	0.002	.0449109	.18528
jiangxi	-.1352577	.0579781	-2.33	0.027	-.2542188	-.0162965
jilin	-.2220282	.2253552	-0.99	0.333	-.6844189	.2403624
liaoning	-.2789811	.172656	-1.62	0.118	-.6332419	.0752797
neimong	-.9288069	.2561317	-3.63	0.001	-1.454346	-.403268
ningxia	-.8813594	.1975659	-4.46	0.000	-1.286731	-.4759877
qinghai	-.7062497	.1521719	-4.64	0.000	-1.018481	-.3940187
shaanxi	-.3342067	.0925991	-3.61	0.001	-.5242045	-.144209
shangdong	-.0049215	.0581511	-0.08	0.933	-.1242377	.1143947
shanghai	.113901	.0648627	1.76	0.090	-.0191862	.2469882
shanxi	-.5312338	.1514863	-3.51	0.002	-.8420581	-.2204095
sichuan	.0251618	.0320732	0.78	0.440	-.040647	.0909707
tianjin	-.3047612	.1190042	-2.56	0.016	-.5489376	-.0605848
xinjiang	-.4007117	.2397817	-1.67	0.106	-.8927032	.0912797
yunnan	-.2774542	.0632713	-4.39	0.000	-.4072763	-.1476321
zhejiang	.1764445	.0822195	2.15	0.041	.007744	.345145
_cons	2.568156	.8279625	3.10	0.004	.8693177	4.266995

不少个体虚拟变量在 5%水平上显著，故可拒绝“所有个体虚拟变量都为 0”的原假设，即认为存在个体固定效应，不应使用混合回归。

LSDV 法的回归系数与组内估计量完全相同，但聚类稳健的标准误略有差别。

对于固定效应模型，也可使用一阶差分法(FD)。Stata 没有专门执行一阶差分法的命令，但在使用命令“xtserial,output”对组内自相关进行检验时，附带提供一阶差分法的估计结果(将此结果记为“FD”)：

```
  . xtserial ltvfo ltlan ltwlab ltpow ltfer hrs
mipric1 giprice mci ngca,output
  . estimates store FD
```

Linear regression

Number of obs	= 420
F(9, 27)	= 902.61
Prob > F	= 0.0000
R-squared	= 0.5797
Root MSE	= .11179

(Std. err. adjusted for 28 clusters in province)

D.ltvfo	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
ltlan D1.	.9807158	.0926143	10.59	0.000	.790687	1.170745
ltwlab D1.	.2420082	.0734117	3.30	0.003	.0913798	.3926366
ltpow D1.	-.0171023	.0747984	-0.23	0.821	-.170576	.1363714
ltfer D1.	.2768317	.0589799	4.69	0.000	.155815	.3978485

```
         hrs |
          D1.|   .2427773   .0372382     6.52   0.000     .1663709    .3191837
             |
     mipric1 |
          D1.|   .0250908   .0357935     0.70   0.489    -.0483513    .0985329
             |
     giprice |
          D1.|  -.0157708    .021774    -0.72   0.475    -.0604473    .0289057
             |
         mci |
          D1.|   .1314675   .1260309     1.04   0.306    -.1271266    .3900616
             |
        ngca |
          D1.|  -.0260777   .4846049    -0.05   0.957    -1.020405    .9682494
------------------------------------------------------------------------------

Wooldridge test for autocorrelation in panel data
H0: no first-order autocorrelation
    F(  1,      27) =     12.511
           Prob > F =      0.0015
```

一阶差分估计量(FD)的估计系数与组内估计量(FE)有一定差别。一般认为，FE 比 FD 更有效率，故较少使用 FD。

也可以在固定效应模型中考虑时间效应，即双向固定效应(Two-way FE)，以捕捉技术进步等效应。

为节省待估参数，首先考虑加入时间趋势项(将估计结果记为“FE_trend”)：

```
. xtreg ltvfo ltlan ltwlab ltpow ltfer hrs mipric1 giprice mci ngca t,fe r

. estimates store FE_trend
```

```
Fixed-effects (within) regression               Number of obs     =         476
Group variable: province                        Number of groups  =          28

R-squared:                                      Obs per group:
     Within  = 0.8749                                         min =          17
     Between = 0.6490                                         avg =        17.0
     Overall = 0.7006                                         max =          17

                                                F(10, 27)         =      247.93
corr(u_i, Xb) = -0.3767                         Prob > F          =      0.0000

                                (Std. err. adjusted for 28 clusters in province)
------------------------------------------------------------------------------
             |               Robust
       ltvfo |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ltlan |   .6517195   .1843858     3.53   0.001     .2733911    1.030048
      ltwlab |   .1431791   .0589267     2.43   0.022     .0222716    .2640866
       ltpow |   .0366317   .0991178     0.37   0.715    -.1667413    .2400047
       ltfer |    .180359   .0428995     4.20   0.000     .0923365    .2683816
         hrs |   .1916276   .0295596     6.48   0.000     .1309763    .2522789
     mipric1 |   .0198772   .0515121     0.39   0.703    -.0858168    .1255713
     giprice |   -.026268   .0226875    -1.16   0.257    -.0728189    .0202829
         mci |   .2014685    .078794     2.56   0.016     .0397965    .3631404
        ngca |   .6761116    .421738     1.60   0.121    -.1892234    1.541447
           t |   .0063068   .0106492     0.59   0.559    -.0155436    .0281572
       _cons |    2.36174   .8262751     2.86   0.008     .6663633    4.057116
-------------+----------------------------------------------------------------
     sigma_u |  .30327958
     sigma_e |  .10589784
         rho |  .89132628   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

时间趋势项 t 并不显著(p 值为 0.559)，主要变量的显著性不变。

其次，加入年度虚拟变量。为演示目的，定义年度虚拟变量：

```
. tab year,gen(year)
```

year	Freq.	Percent	Cum.
70	28	5.56	5.56
71	28	5.56	11.11
72	28	5.56	16.67
73	28	5.56	22.22
74	28	5.56	27.78
75	28	5.56	33.33
76	28	5.56	38.89
77	28	5.56	44.44
78	28	5.56	50.00
79	28	5.56	55.56
80	28	5.56	61.11
81	28	5.56	66.67
82	28	5.56	72.22
83	28	5.56	77.78
84	28	5.56	83.33
85	28	5.56	88.89
86	28	5.56	94.44
87	28	5.56	100.00
Total	504	100.00	

此命令将生成时间虚拟变量 *year1*，*year2*，…，*year18*。

加入年度虚拟变量后，由于两个价格变量 *mipric1* 与 *giprice* 在各省都一样，故无法包括在回归方程中，以避免严格多重共线性。

进行含时间虚拟变量的双向固定效应估计(将结果记为“FE_TW”)：

```
. xtreg ltvfo ltlan ltwlab ltpow ltfer hrs mci
ngca year2-year18,fe r

. estimates store FE_TW
```

Fixed-effects (within) regression — Number of obs = 476
Group variable: province — Number of groups = 28

R-squared:
Within = 0.8932
Between = 0.6596
Overall = 0.7156

Obs per group:
min = 17
avg = 17.0
max = 17

F(23, 27) = 949.82
corr(u_i, Xb) = -0.3425 — Prob > F = 0.0000

(Std. err. adjusted for 28 clusters in province)

ltvfo	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
ltlan	.5833594	.1745834	3.34	0.002	.2251439	.9415749
ltwlab	.1514909	.0585107	2.59	0.015	.0314368	.271545
ltpow	.0971114	.090911	1.07	0.295	-.0894225	.2836453
ltfer	.1693346	.0438098	3.87	0.001	.0794444	.2592248
hrs	.1503752	.0587581	2.56	0.016	.0298136	.2709368
mci	.1978373	.0810587	2.44	0.022	.0315186	.364156
ngca	.7784081	.4016301	1.94	0.063	-.0456688	1.602485
year2	-.0240404	.023366	-1.03	0.313	-.0719836	.0239027
year3	-.1323624	.0404832	-3.27	0.003	-.2154272	-.0492977

year4	-.0377336	.0357883	-1.05	0.301	-.111165	.0356979
year5	.0058554	.0500774	0.12	0.908	-.096895	.1086058
year6	.0096731	.0566898	0.17	0.866	-.1066448	.1259911
year7	-.0476465	.061423	-0.78	0.445	-.1736761	.0783832
year8	-.0869336	.0680579	-1.28	0.212	-.2265767	.0527096
year9	-.0325205	.0766428	-0.42	0.675	-.1897785	.1247376
year10	-.0076332	.0833462	-0.09	0.928	-.1786454	.163379
year11	0	(omitted)				
year12	-.093479	.1093614	-0.85	0.400	-.3178701	.1309121
year13	-.0447862	.1207405	-0.37	0.714	-.2925251	.2029528
year14	-.0309435	.1377207	-0.22	0.824	-.313523	.2516361
year15	.0442535	.1428764	0.31	0.759	-.2489048	.3374117
year16	-.0033372	.1561209	-0.02	0.983	-.3236709	.3169965
year17	.00484	.157992	0.03	0.976	-.3193329	.3290129
year18	.0386475	.1639608	0.24	0.815	-.2977723	.3750674
_cons	2.651286	.7738994	3.43	0.002	1.063376	4.239196
sigma_u	.29344594					
sigma_e	.09930555					
rho	.89724523	(fraction of variance due to u_i)				

year1(即 1970 年)被作为基期(对应于常数项_cons)，而不包括在上述回归命令中。

由于 1980 年的 hrs 数据缺失，故 *year11*(即 1980 年)也被去掉。

即使在双向固定效应模型中，*hrs* 也依然在 5%水平上显著为正。

大多数的年度虚拟变量均不显著(但 *year3* 在 1%水平上显著)。

检验所有年度虚拟变量的联合显著性：

```
. test year2 year3 year4 year5 year6 year7 year8 year9 year10 year12 year13  year14 year15 year16 year17 year18
```

```
( 1)  year2 = 0
( 2)  year3 = 0
( 3)  year4 = 0
( 4)  year5 = 0
( 5)  year6 = 0
( 6)  year7 = 0
( 7)  year8 = 0
( 8)  year9 = 0
( 9)  year10 = 0
(10)  year12 = 0
(11)  year13 = 0
(12)  year14 = 0
(13)  year15 = 0
(14)  year16 = 0
(15)  year17 = 0
(16)  year18 = 0

      F( 16,    27) =   14.82
           Prob > F =    0.0000
```

结果强烈拒绝“无时间固定效应”的原假设，认为应在模型中包括时间固定效应。

还可直接用以下命令来估计双向固定效应模型(不必先生成时间虚拟变量)：

```
. xtreg ltvfo ltlan ltwlab ltpow ltfer hrs mci ngca i.year,fe r
```

其中，“i.year”表示根据变量 year 的不同取值来生成年度虚拟变量。

```
Fixed-effects (within) regression               Number of obs     =         476
Group variable: province                        Number of groups  =          28

R-squared:                                      Obs per group:
     Within  = 0.8932                                         min =          17
     Between = 0.6596                                         avg =        17.0
     Overall = 0.7156                                         max =          17

                                                F(23, 27)         =      949.82
corr(u_i, Xb) = -0.3425                         Prob > F          =      0.0000

                                (Std. err. adjusted for 28 clusters in province)
```

ltvfo	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
ltlan	.5833594	.1745834	3.34	0.002	.2251439	.9415749
ltwlab	.1514909	.0585107	2.59	0.015	.0314368	.271545
ltpow	.0971114	.090911	1.07	0.295	-.0894225	.2836453
ltfer	.1693346	.0438098	3.87	0.001	.0794444	.2592248
hrs	.1503752	.0587581	2.56	0.016	.0298136	.2709368
mci	.1978373	.0810587	2.44	0.022	.0315186	.364156
ngca	.7784081	.4016301	1.94	0.063	-.0456688	1.602485

year						
71	-.0240404	.023366	-1.03	0.313	-.0719836	.0239027
72	-.1323624	.0404832	-3.27	0.003	-.2154272	-.0492977
73	-.0377336	.0357883	-1.05	0.301	-.111165	.0356979
74	.0058554	.0500774	0.12	0.908	-.096895	.1086058
75	.0096731	.0566898	0.17	0.866	-.1066448	.1259911
76	-.0476465	.061423	-0.78	0.445	-.1736761	.0783832
77	-.0869336	.0680579	-1.28	0.212	-.2265767	.0527096
78	-.0325205	.0766428	-0.42	0.675	-.1897785	.1247376
79	-.0076332	.0833462	-0.09	0.928	-.1786454	.163379
81	-.093479	.1093614	-0.85	0.400	-.3178701	.1309121
82	-.0447862	.1207405	-0.37	0.714	-.2925251	.2029528
83	-.0309435	.1377207	-0.22	0.824	-.313523	.2516361
84	.0442535	.1428764	0.31	0.759	-.2489048	.3374117
85	-.0033372	.1561209	-0.02	0.983	-.3236709	.3169965
86	.00484	.157992	0.03	0.976	-.3193329	.3290129
87	.0386475	.1639608	0.24	0.815	-.2977723	.3750674
_cons	2.651286	.7738994	3.43	0.002	1.063376	4.239196
sigma_u	.29344594					
sigma_e	.09930555					
rho	.89724523	(fraction of variance due to u_i)				

4. 随机效应

以上结果已基本确认了个体效应的存在，但个体效应仍可能以随机效应(RE)的形式存在。

随机效应估计的 Stata 命令为

```
. xtreg y x1 x2 x3,re r theta
```

选择项“re”为默认选项(可省略)

选择项“r”表示使用聚类稳健标准误，如果使用选择项“vce(cluster id)”也能达到完全相同的效果。

选择项“theta”表示显示用于进行广义离差变换的θ值。

对于随机效应模型，也可以进行 MLE 估计，其 Stata 命令为

```
. xtreg y x1 x2 x3,mle
```

下面，进行随机效应(RE)的估计(将结果记为“RE_robust”)：

```
. xtreg ltvfo ltlan ltwlab ltpow ltfer hrs mci ngca,re r theta

. estimates store RE_robust
```

```
Random-effects GLS regression                   Number of obs     =         476
Group variable: province                        Number of groups  =          28

R-squared:                                      Obs per group:
     Within  = 0.8700                                         min =          17
     Between = 0.8135                                         avg =        17.0
     Overall = 0.8263                                         max =          17

                                                Wald chi2(7)      =     2452.50
corr(u_i, X) = 0 (assumed)                      Prob > chi2       =      0.0000
theta        = .81012778

                                (Std. err. adjusted for 28 clusters in province)
------------------------------------------------------------------------------
             |               Robust
       ltvfo |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ltlan |   .5655915   .1089863     5.19   0.000     .3519823    .7792007
      ltwlab |   .1441844   .0462225     3.12   0.002     .0535899     .234779
       ltpow |    .060477   .0508828     1.19   0.235    -.0392515    .1602055
       ltfer |   .1882741   .0386418     4.87   0.000     .1125376    .2640107
         hrs |   .2186096   .0377121     5.80   0.000     .1446952    .2925241
         mci |   .4702368   .0836862     5.62   0.000      .306215    .6342587
        ngca |   .6745175   .3663329     1.84   0.066    -.0434818    1.392517
       _cons |   2.387878   .5672669     4.21   0.000     1.276055    3.499701
-------------+----------------------------------------------------------------
     sigma_u |  .13324845
     sigma_e |  .10624809
         rho |   .6113231   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

上表显示，$\sigma_u = 0.13324845$，$\sigma_\varepsilon = 0.10624809$，而$\rho \equiv \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2} =$ 0.6113231。

究竟应使用混合回归，还是个体随机效应模型？

Breusch and Pagan (1980)提供了一个检验个体随机效应的LM检验，其原假设为“$H_0 : \sigma_u^2 = 0$”，而备择假设为“$H_1 : \sigma_u^2 \neq 0$”。

如果拒绝H_0，则说明原模型中应该有一个反映个体特性的随机扰动项u_i，而不应该使用混合回归。

该 *LM* 检验的 Stata 命令为“xttest0”(在执行命令“xtreg,re”之后才能进行)。

```
. xttest0
```

```
Breusch and Pagan Lagrangian multiplier test for random effects

        ltvfo[province,t] = Xb + u[province] + e[province,t]

        Estimated results:
                         |       Var     SD = sqrt(Var)
                 ---------+------------------------------
                    ltvfo |    .2893078       .5378734
                        e |    .0112887       .1062481
                        u |    .0177551       .1332484

        Test: Var(u) = 0
                                  chibar2(01) =  1235.75
                              Prob > chibar2 =   0.0000
```

LM 检验强烈拒绝“不存在个体随机效应”的原假设。

看一下使用普通标准误的随机效应估计结果(将结果记为“RE”)。

```
. xtreg ltvfo ltlan ltwlab ltpow ltfer hrs mci ngca,re

. estimates store RE
```

```
Random-effects GLS regression                   Number of obs     =       476
Group variable: province                        Number of groups  =        28

R-squared:                                      Obs per group:
     Within  = 0.8700                                         min =        17
     Between = 0.8135                                         avg =      17.0
     Overall = 0.8263                                         max =        17

                                                Wald chi2(7)      =   2981.73
corr(u_i, X) = 0 (assumed)                      Prob > chi2       =    0.0000
```

ltvfo	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
ltlan	.5655915	.0478214	11.83	0.000	.4718633	.6593196
ltwlab	.1441844	.0233398	6.18	0.000	.0984394	.1899295
ltpow	.060477	.0252917	2.39	0.017	.0109062	.1100478
ltfer	.1882741	.0208337	9.04	0.000	.1474408	.2291075
hrs	.2186096	.0216932	10.08	0.000	.1760918	.2611275
mci	.4702368	.064681	7.27	0.000	.3434643	.5970093
ngca	.6745175	.2121571	3.18	0.001	.2586973	1.090338
_cons	2.387878	.2895274	8.25	0.000	1.820414	2.955341
sigma_u	.13324845					
sigma_e	.10624809					
rho	.6113231	(fraction of variance due to u_i)				

作为对照，也可以对随机效应模型进行 MLE 估计(将结果记为“MLE”)：

```
. xtreg ltvfo ltlan ltwlab ltpow ltfer hrs mci ngca,mle nolog

. estimates store MLE
```

```
Random-effects ML regression                    Number of obs    =    476
Group variable: province                        Number of groups =     28

Random effects u_i ~ Gaussian                   Obs per group:
                                                              min =     17
                                                              avg =   17.0
                                                              max =     17

                                                LR chi2(7)       = 961.00
Log likelihood = 332.89739                      Prob > chi2      = 0.0000
```

ltvfo	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
ltlan	.5643288	.0518396	10.89	0.000	.462725	.6659326
ltwlab	.1408974	.0235161	5.99	0.000	.0948067	.1869882
ltpow	.0717013	.0248585	2.88	0.004	.0229794	.1204231
ltfer	.1772027	.0206075	8.60	0.000	.1368127	.2175927
hrs	.2153264	.020987	10.26	0.000	.1741926	.2564602
mci	.4064832	.0712859	5.70	0.000	.2667654	.546201
ngca	.7149811	.2105466	3.40	0.001	.3023172	1.127645
_cons	2.490512	.3039688	8.19	0.000	1.894744	3.08628
/sigma_u	.2141094	.0317827			.1600597	.2864107
/sigma_e	.1061021	.0035665			.0993371	.1133278
rho	.8028448	.0486812			.6942946	.8840722

```
LR test of sigma_u=0: chibar2(01) = 464.22              Prob >= chibar2 = 0.000
```

上表显示，随机效应 MLE 的系数估计值与随机效应 FGLS 有所不同，但在性质上依然类似。

上表最后一行的LR检验强烈拒绝原假设“$H_0:\sigma_u=0$”，即认为存在个体随机效应，不应进行混合回归。

5．固定效应还是随机效应：豪斯曼检验

在处理面板数据时，究竟使用固定效应还是随机效应模型，这是一个基本问题。

为此，需进行豪斯曼检验。

豪斯曼检验的 Stata 命令为

```
. xtreg y x1 x2 x3,fe        (固定效应估计)
. estimates store FE        (存储结果)
. xtreg y x1 x2 x3,re        (随机效应估计)
. estimates store RE        (存储结果)
. hausman FE RE,constant sigmamore  (豪斯曼检验)
```

选择项“constant”表示在比较系数估计值时包括常数项(默认不含常数项)；

选择项“sigmamore”表示统一使用更有效率的那个估计量(即随机效应估计量)的方差估计。

由于传统的豪斯曼检验假设球形扰动项，故在进行固定效应与随机效应的估计时，均不使用异方差或聚类稳健的标准误。

由于前面已存储了相应的估计结果，故可直接进行豪斯曼检验。

```
. hausman FE RE,constant sigmamore
```

	Coefficients (b) FE	(B) RE	(b-B) Difference	sqrt(diag(V_b-V_B)) Std. err.
ltlan	.6370234	.5655915	.0714319	.0522174
ltwlab	.1387786	.1441844	-.0054059	.0145627
ltpow	.0577152	.060477	-.0027618	.0241549
ltfer	.1826281	.1882741	-.005646	.0100505
hrs	.2134022	.2186096	-.0052075	.0091613
mci	.1943697	.4702368	-.2758671	.0657487
ngca	.7562031	.6745175	.0816856	.083631
_cons	2.337895	2.387878	-.0499825	.2834941

b = Consistent under H0 and Ha; obtained from **xtreg**.
B = Inconsistent under Ha, efficient under H0; obtained from **xtreg**.

Test of H0: Difference in coefficients not systematic

chi2(8) = (b-B)'[(V_b-V_B)^(-1)](b-B)
= 38.56
Prob > chi2 = 0.0000
(V_b-V_B is not positive definite)

由于p值为0.0000，故强烈拒绝原假设“$H_0:u_i$与解释变量不相关”，认为应该使用固定效应模型，而非随机效应模型。

传统的豪斯曼检验假定，在H_0成立的情况下，随机效应模型最有效率。

这意味着，扰动项必须是同方差的，在异方差的情况下并不适用。

在 Stata 中进行以上豪斯曼检验时，如果使用聚类稳健标准误，比如“`xtreg y x1 x2 x3,fe vce(cluster id)`”，则 Stata 将无法执行“`hausman FE RE`”命令。

为此，下载非官方命令 `xtoverid` 进行稳健的豪斯曼检验。

此处“`overid`”指“overidentification test”(过度识别检验)。

因为随机效应模型与固定效应模型相比，前者多了“个体异质性u_i与解释变量不相关”的约束条件，也可视为过度识别条件。

```
. ssc install xtoverid          (下载安装命令 xtoverid)
```

在使用命令 xtoverid 之前，须先以稳健标准误来执行命令“xtreg,re”。

```
. quietly xtreg ltvfo ltlan ltwlab ltpow ltfer hrs mci ngca,r

. xtoverid
```

```
Test of overidentifying restrictions: fixed vs random effects
Cross-section time-series model: xtreg re  robust cluster(province)
Sargan-Hansen statistic 221.225  Chi-sq(7)    P-value = 0.0000
```

$\chi^2(7)$统计量为 221.225，p值为 0.0000，故仍然强烈拒绝“随机效应”的原假设。

6. 组间估计量

纯粹为了演示目的，下面进行组间估计。

```
. xtreg ltvfo ltlan ltwlab ltpow ltfer hrs mci ngca,be
```

```
Between regression (regression on group means)  Number of obs     =         476
Group variable: province                        Number of groups  =          28

R-squared:                                      Obs per group:
     Within  = 0.4673                                         min =          17
     Between = 0.9362                                         avg =        17.0
     Overall = 0.0232                                         max =          17

                                                F(7,20)           =       41.93
sd(u_i + avg(e_i.)) = .1357173                  Prob > F          =      0.0000
```

ltvfo	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
ltlan	.7021718	.1177241	5.96	0.000	.4566037	.9477398
ltwlab	.5020747	.0940029	5.34	0.000	.3059881	.6981612
ltpow	-.1511518	.1188815	-1.27	0.218	-.3991344	.0968307
ltfer	.1132485	.0976941	1.16	0.260	-.0905377	.3170347
hrs	-3.757418	1.242961	-3.02	0.007	-6.350189	-1.164647
mci	.586029	.1838616	3.19	0.005	.2025005	.9695575
ngca	.4443814	.6565795	0.68	0.506	-.9252194	1.813982
_cons	2.432839	1.005985	2.42	0.025	.3343904	4.531288

由于豪斯曼检验选择了固定效应，而组间估计量仅在随机效应成立的情况下才一致，故组间估计的结果并不可信。

例如，根据组间估计量，*hrs* 在 1%水平上显著为负，即家庭联产承包责任制反而对种植业产值有负作用。

下面，将以上各主要方法的回归系数及标准误列表(为节省空间，不汇报包含年度虚拟变量的双向固定效应)：

```
. esttab OLS FE_robust FE_trend FE RE,b se mtitle
```

	(1) OLS	(2) FE_robust	(3) FE_trend	(4) FE	(5) RE
ltlan	0.694***	0.637***	0.652**	0.637***	0.566***
	(0.115)	(0.168)	(0.184)	(0.0673)	(0.0478)
ltwlab	0.265***	0.139*	0.143*	0.139***	0.144***
	(0.0566)	(0.0625)	(0.0589)	(0.0262)	(0.0233)
ltpow	-0.0292	0.0577	0.0366	0.0577	0.0605*
	(0.0670)	(0.0756)	(0.0991)	(0.0333)	(0.0253)
ltfer	0.311***	0.183***	0.180***	0.183***	0.188***
	(0.0531)	(0.0436)	(0.0429)	(0.0220)	(0.0208)
hrs	0.229***	0.213***	0.192***	0.213***	0.219***
	(0.0489)	(0.0391)	(0.0296)	(0.0224)	(0.0217)
mipric1	0.0122	0.0544	0.0199	0.0544	
	(0.0548)	(0.0590)	(0.0515)	(0.0422)	
giprice	-0.0539	-0.0151	-0.0263	-0.0151	
	(0.0274)	(0.0246)	(0.0227)	(0.0187)	
mci	0.695***	0.194*	0.201*	0.194*	0.470***
	(0.169)	(0.0771)	(0.0788)	(0.0877)	(0.0647)
ngca	0.305	0.756	0.676	0.756***	0.675**
	(0.522)	(0.382)	(0.422)	(0.217)	(0.212)
t			0.00631		
			(0.0106)		
_cons	1.081	2.338*	2.362**	2.338***	2.388***
	(0.827)	(0.855)	(0.826)	(0.385)	(0.290)
N	476	476	476	476	476

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001