МИНИСТЕРСТВО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ЭКОНОМИКИ, СТАТИСТИКИ И ИНФОРМАТИКИ

Кафедра "Математической статистики и эконометрики"

В.А.Балаш, О.С. Балаш

Модели линейной регрессии для панельных данных

Учебное пособие

Москва 2002

МИНИСТЕРСТВО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ЭКОНОМИКИ, СТАТИСТИКИ И ИНФОРМАТИКИ

Кафедра "Математической статистики и эконометрики"

В.А.Балаш, О.С. Балаш

МОДЕЛИ ЛИНЕЙНОЙ РЕГРЕССИИ ДЛЯ ПАНЕЛЬНЫХ ДАННЫХ

Рекомендовано Учебно-методическим объединением
по статистике и математическим методам в экономике в качестве учебного пособия
для экономических специальностей вузов

Москва - 2002

УДК

В пособии излагаются методы расчета оценок коэффициентов линейных регрессионных моделей и проверки гипотез для панельных данных.

Предназначено для студентов и аспирантов, обучающихся по специальности «Статистика».

ВВЕДЕНИЕ

Панельными называют данные, содержащие сведения об одном и том же множестве объектов за ряд последовательных периодов времени. Термин "панельный метод" произошел от английского слова panel. В качестве панели могут выступать индивидуумы, группы лиц, предприятия, домохозяйства, регионы, страны и т.д., сведения о которых собраны в течение нескольких периодов времени. Этот метод используют при изучении потребительского поведения, занятости, безработицы, доходов и заработной платы, производственных функций и политики дивидендов фирм, в международных и межрегиональных сопоставлениях. Методы анализа панельных данных могут применяться практически во всех отраслях социально-экономической статистики, так как статистическая отчетность характеризует одни и те же объекты - предприятия, отрасли, регионы с помощью единой системы показателей, которые регулярно фиксируются в заданные периоды или моменты времени. Простое объединение данных за разные годы (метод "заводо-лет") и применение к ним стандартных математико-статистических методов не всегда оправдано. Его использование обосновалось стабильными условиями деятельности предприятий в условиях плановой экономики. Но в условиях резких изменений экономической конъюнктуры, масштаба и структуры цен данные становятся несопоставимыми, поэтому необходимо использовать модели, учитывающие эти особенности.

1. ОСОБЕННОСТИ ПАНЕЛЬНЫХ ДАННЫХ

1.1. Структура панельных данных

Традиционно выборочные данные представляют в виде таблиц «объект-признак»: по строкам располагают объекты, по столбцам – признаки. Для панельных данных добавляется еще одно измерение – время. Например, совокупность данных состоит из 1000 наблюдений, из них первые 100 – значения некоторых признаков для 100 человек в 1990 году, следующие 100 – тех же самых людей за 1991 год и так далее до 1999 года.

Панельные данные можно также представить в виде таблицы «объект-признак». При этом придерживаются следующего соглашения. Признаки располагаются по столбцам, по строкам – данные о первом объекте за T периодов (строки 1, 2,3,…,T), затем о втором объекте (строки $T+1$, $T+2$,…$2T$) и т.д. Всего NT строк.

Объекты		**Признаки**		
	$t=1$	X_{11}	Y_{11}	Z_{11} ……
объект 1	$t=2$	X_{12}	Y_{12}	Z_{12}……
	….	…….	……	……
	$t=T$	X_{1T}	Y_{1T}	Z_{1T} ……
	$t=1$	X_{21}	Y_{21}	Z_{21} ……
объект 2	$t=2$	X_{22}	Y_{22}	Z_{22}……
	….	…….	……	……
	$t=T$	X_{2T}	Y_{2T}	Z_{2T} ……
……	……	……	……	……
	$t=1$	X_{N1}	Y_{N1}	Z_{N1} ……
объект N	$t=2$	X_{N2}	Y_{N2}	Z_{N2}……
	….	……	……	……
	$t=T$	X_{NT}	Y_{NT}	Z_{NT} ……

Панель бывает сбалансированная и несбалансированная. Если данные присутствуют по всем объектам за все периоды времени, то панель называется ***сбалансированной*** (рис.1).

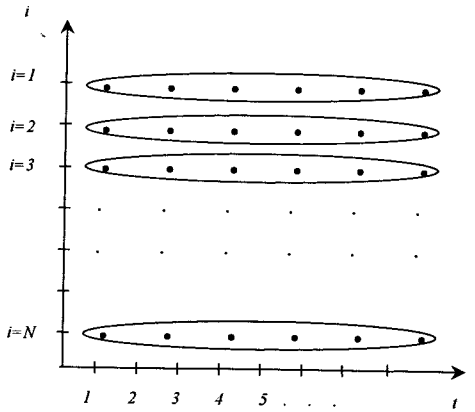

Рис. 1. Сбалансированная панель

Достаточно часто из-за технических, организационных или иных причин в некоторые периоды времени не удается собрать сведения для всех объектов, включенных в выборку первоначально (смерть, отъезд, болезнь и т.п.). Чтобы сохранить репрезентативность, отсутствующие объекты приходится заменять другими. В результате получим ***несбалансированную панель*** (рис. 2).

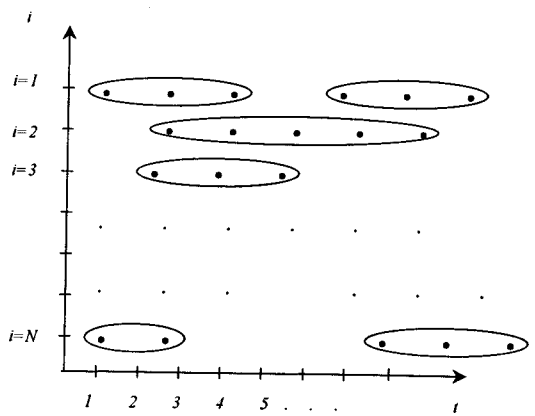

Рис. 2. Несбалансированная панель

При исследовании проблем занятости и безработицы в международной практике распространены так называемые ***ротационные панели***. Объект (человек трудоспособного возраста) участвует в шести последовательных ежеквартальных опросах, а затем исключается из панели. Таким образом, 1/6 часть всей выборки обновляется (рис.3).

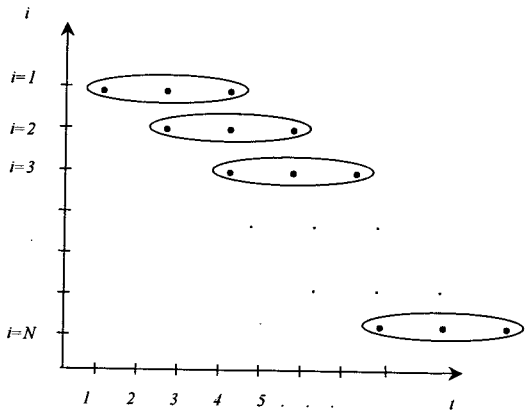

Рис. 3. Ротационная панель

Возможны и иные модификации панельных данных. Но наибольшее распространение получили сбалансированные и несбалансированные панели.

В большинстве случаев число объектов достаточно велико (несколько десятков, сотен или тысяч), а число моментов наблюдения ограничено. Наиболее длинный временной ряд панельного исследования динамики доходов (PSID), проводимые Мичиганским университетом содержат данные о 6000 семей США и 15000 респондентах с 1968 года по настоящее время. В других случаях временные ряды существенно короче.

В России наибольшей известностью пользуется панельное исследование «Российский мониторинг экономики и здоровья населения» (Russian Longitudinal Monitoring Survey - RLMS), которое охватывает информацию по более чем 3000 семей за 1995-2000 годы.

Преимущества панельных данных следующие. Во-первых, большее число наблюдений (*NT*) обеспечивает большую эффективность оценивания параметров эконометрической модели. Во-вторых, появляется возможность контроля над неоднородностью объектов. В-третьих, возможностью идентифицировать эффекты, недоступные в анализе кросс-данных.

1.2. Скрытые переменные и индивидуальные эффекты

Прежде всего, обратим внимание на то, что использование панельных данных позволяет более полно учесть особенности объектов, попавших в выборку. Каждый человек или хозяйствующий субъект обладает некоторыми признаками, которые могут воздействовать на результативный показатель, но плохо поддаются регистрации, то есть являются неучтенными, скрытыми или ненаблюдаемыми. Если их значения различны для разных объектов, но постоянны во времени, их влияние можно учесть, вводя в модель индивидуальные уровни для каждого объекта.

Рассмотрим несколько примеров.

Пусть задача исследования заключается в оценке отдачи от образования - влияния числа лет обучения на почасовую заработную плату. Предположим, что почасовая заработная плата определяется двумя факторами – "образованием" и "способностями".

Существует несколько способов сформировать панельный массив данных. Первый – во времени, опрашивая несколько лет подряд одну и ту же группу людей. Но есть интересная альтернатива, состоящая в том, что можно попытаться отыскать достаточное количество пар близнецов, провести единовременный опрос и сравнить их доходы в текущем году. Так как близнецы генетически идентичны, можно предположить, что способности у них одинаковы. Тогда в качестве объекта выступает пара близнецов, а аналога «времени» - номер близнеца в паре.

Пусть $k = 1,2,...K$ – соответствует номеру пары, $j=1,2$ – номеру близнеца. Тогда общее число наблюдений равно $N=2K$.

Модель зависимости текущей заработной платы от образования имеет вид:

$$\ln(Y_{jk}) = \beta_0 + \beta_1 S_{jk} + \beta_2 A_{jk} + \varepsilon_{jk},$$

где Y_{jk} – заработная плата, S_{jk} – уровень образования (число лет обучения), A_{jk} – "способности" индивида.

Так как по предположению, способности у близнецов совпадают, то для любого k:

$$A_{1k} = A_{2k} = A_k.$$

Хотя способности не поддаются непосредственному измерению, естественно допустить, что более талантливые люди получают лучшее образование:

$$\mathrm{cov}(S_{jk}, A_k) > 0.$$

Допустим, что в некоторых парах близнецы по каким-либо причинам получили разное образование.

Тогда модель можно переписать в виде:

$$\ln(Y_{jk}) = \beta_0 + \beta_1 S_{jk} + \beta_2 A_k + \varepsilon_{jk}.$$

Заметим, что способности индивида не поддаются прямому измерению, то есть мы не располагаем сведениями о величинах A_k. Так как уравнение включает ненаблюдаемую переменную, мы не можем оценить все коэффициенты модели.

Наиболее простой вариант состоит в том, чтобы вообще исключить переменную A_k из рассмотрения:

$$\ln(Y_{jk}) = \beta_0 + \beta_1 S_{jk} + \varepsilon_{jk}.$$

Условные данные для 10 пар близнецов приведены в таблице 1.

Таблица 1

Зависимость почасовой оплаты от числа лет учебы для 10 пар близнецов

Номер пары *k*	Номер индивида в паре *j*	Число лет учебы *S*	Почасовая оплата *Y*, ден.ед.
1	1	10	11,3
1	2	12	11,2
2	1	11	13,3
2	2	16	19,2
3	1	10	13,7
3	2	10	12,6
4	1	15	27,8
4	2	15	19,6
5	1	12	15,8

Номер пары *k*	Номер индивида в паре *j*	Число лет учебы *S*	Почасовая оплата *Y*, ден.ед.
5	2	14	22,0
6	1	10	16,0
6	2	12	13,5
7	1	11	22,8
7	2	15	19,8
8	1	15	21,5
8	2	15	27,8
9	1	11	17,5
9	2	14	18,8
10	1	14	25,1
10	2	15	30,3

Применяя метод наименьших квадратов для данных таблицы 1, получим следующие оценки коэффициентов:

$$\ln(\hat{Y}_{jk}) = 1{,}5613 + 0{,}1041 S_{jk}.$$

Из полученного уравнения следует, что повышение уровня образования на один год приводит в среднем к повышению почасовой оплаты на 10,41 процента. Эта оценка будет смещена, скорее всего завышена, из-за того, что неучтенная переменная "способности" (A_k) сильно коррелированна с объясняющей переменной "образование" (S_{jk}).

Использование панельных данных позволяет устранить это смещение.

Введем K фиктивных переменных:

$$D_{jk}^{k=1} = \begin{cases} 1, \text{для первой пары близнецов,} \\ 0, \text{для остальных пар.} \end{cases}$$

$$D_{jk}^{k=2} = \begin{cases} 1, \text{для второй пары близнецов,} \\ 0, \text{для остальных пар.} \end{cases}$$

....

$$D_{jk}^{k=K} = \begin{cases} 1, \text{для } K-\text{ой пары близнецов,} \\ 0, \text{для остальных пар.} \end{cases}$$

В этом случае:

$$A_k = A_1 D_{jk}^{k=1} + A_2 D_{jk}^{k=2} + ... + A_K D_{jk}^{k=K}.$$

Если включить в уравнение регрессии все K фиктивных переменных и свободный член β_0, будет строгая мультиколлениарность.

Поэтому один из параметров следует исключить из модели, например, свободный член β_0.

Перепишем уравнение регрессии без свободного члена:

$$\ln(Y_{jk}) = \beta_1 S_{jk} + \beta_2 (A_1 D_{jk}^{k=1} + A_2 D_{jk}^{k=2} + ... + A_K D_{jk}^{k=K}) + \varepsilon_{jk}.$$

Получим модель с $K+1$ независимой переменной:

$$\ln(Y_{jk}) = \beta_1 S_{jk} + (\beta_2 A_1) D_{jk}^{k=1} + (\beta_2 A_2) D_{jk}^{k=2} + ... + (\beta_2 A_K) D_{jk}^{k=K} + \varepsilon_{jk}.$$

Обозначим $\delta_i = \beta_2 A_i$, $i=1, 2, ...K$. Тогда неизвестные величины A_i включаются в уравнение как часть параметров δ_i, подлежащих оценке:

$$\ln(Y_{jk}) = \beta_1 S_{jk} + \delta_1 D_{jk}^{k=1} + \delta_2 D_{jk}^{k=2} + ... + \delta_K D_{jk}^{k=K} + \varepsilon_{jk}.$$

В результате модель не содержит неизмеряемых переменных. Оценив коэффициенты модели методом наименьших квадратов, получим несмещенную оценку параметра β_1.

Коэффициент при каждой фиктивной переменной D_{jk} соответствует влиянию способностей каждой отдельной пары близнецов на их доход. А его произведение на фиктивную переменную - индивидуальный эффект, соответствующей i –ой паре:

$$\mu_i = \delta_i D_{jk}^{k=i} = \beta_2 A_i D_{jk}^{k=i}.$$

Поэтому предыдущее уравнение эквивалентно введению в модель K индивидуальных эффектов:

$$\ln(Y_{jk}) = \mu_k + \beta_1 S_{jk} + \varepsilon_{jk}.$$

Данный подход получил название ***модели с фиксированными эффектами (fixed effects model)***. Использование панельных данных позволило ввести в модель индивидуальные эффекты для того, чтобы избавиться от влияния ненаблюдаемой переменной (постоянной во времени) и получить несмещенную оценку интересующего нас параметра.

Заметим, что если мы назвали бы ненаблюдаемую переменную A_k не «способностями» индивида, а как либо иначе, наши рассуждения не изменились. Важно то, что такая переменная существует и не меняется во времени. Если бы таких переменных было бы не одна, а несколько, влияние их всех аккумулировалось бы в индивидуальных эффектах.

Для данных нашего примера получим:

$$\ln(\Delta \hat{Y}_{jt}) = 0{,}03141 S_{jt} + 2{,}07\, D_1 + 2{,}35\, D_2 + 2{,}26\, D_3 + 2{,}68\, D_4 + 2{,}52\, D_5 + 2{,}34\, D_6 + 2{,}65\, D_7 + 2{,}73\, D_8 + 2{,}51\, D_9 + 2{,}86\, D_{10}$$

Введение фиктивных переменных – вовсе не единственная возможность. Того же результата можно добиться и иным путем. Запишем исходную модель для каждой пары близнецов:

$$\ln(Y_{1k}) = \beta_0 + \beta_1 S_{1k} + \beta_2 A_k + \varepsilon_{1k},$$

$$\ln(Y_{2k}) = \beta_0 + \beta_1 S_{2k} + \beta_2 A_k + \varepsilon_{2k}.$$

Найдем разность между ними:

$$\ln(Y_{1k}) - \ln(Y_{2k}) = \beta_0 + \beta_1 S_{1k} + \beta_2 A_k + \varepsilon_{1k} - (\beta_0 + \beta_1 S_{2k} + \beta_2 A_k + \varepsilon_{2k}),$$

$$\ln(Y_{1k}) - \ln(Y_{2k}) = (\beta_0 - \beta_0) + \beta_1 (S_{1k} - S_{2k}) + \beta_2 (A_k - A_k) + (\varepsilon_{1k} - \varepsilon_{2k}),$$

$$\ln(Y_{1k}) - \ln(Y_{2k}) = \beta_1 (S_{1k} - S_{2k}) + (\varepsilon_{1k} - \varepsilon_{2k}).$$

Полученное уравнение содержит лишь одну объясняющую переменную (S) – уровень образования.

Для преобразованного уравнения можно получить несмещенную оценку интересующего нас параметра β_1 методом наименьших квадратов. Этот прием оценивания обычно называют ***переходом к первым разностям (first differences).***

В общем виде получили:

$$\Delta \ln(Y_{1k}) = \beta_1 \Delta S_{1k} + \Delta \varepsilon_{1k}.$$

Найдем МНК-оценку:

$$b_1 = \frac{\sum \Delta \ln(Y_{1k}) \Delta S_{1k}}{\sum \Delta S_{1k}^2},$$

$$b_1 = \frac{\sum (\ln(Y_{1k}) - \ln(Y_{2k}))(S_{1k} - S_{2k})}{\sum (S_{1k} - S_{2k})^2},$$

$$b_1 = \frac{\sum (\beta_1 (S_{1k} - S_{2k}) + (\varepsilon_{1k} - \varepsilon_{2k}))(S_{1k} - S_{2k})}{\sum (S_{1k} - S_{2k})^2},$$

$$b_1 = \beta_1 + \frac{\sum (\varepsilon_{1k} - \varepsilon_{2k})(S_{1k} - S_{2k})}{\sum (S_{1k} - S_{2k})^2}.$$

Найдем математическое ожидание оценки b_1:

$$E(b_1) = \beta_1 + E\left(\frac{\sum (\varepsilon_{1k} - \varepsilon_{2k})(S_{1k} - S_{2k})}{\sum (S_{1k} - S_{2k})^2}\right),$$

$$E(b_1) = \beta_1 + \frac{\sum E\frac{1}{K}(\varepsilon_{1k} - \varepsilon_{2k})(S_{1k} - S_{2k})}{\sum \frac{1}{K}(S_{1k} - S_{2k})^2} = \beta_1 + \frac{0}{\mathrm{var}(S_{1k} - S_{2k})} = \beta_1.$$

Следовательно, оценка несмещена.

Таким образом, использование панельных данных позволяет элиминировать эффект ненаблюдаемых переменных и получить несмещенную оценку отдачи от образования.

Для данных таблицы 1 после перехода к разностям получим следующую оценку:

$$\Delta \ln(\hat{Y}_{1k}) = 0{,}0314 \Delta S_{1k}.$$

Коэффициент модели b_1=0,0314 показывает, что для рассматриваемого набора данных норма отдачи от образования составляет около 3 процентов. Учет влияния скрытого признака – способностей индивида, снижает норму отдачи от образования почти в три раза - с 10 процентов до 3 процентов на каждый дополнительный год обучения.

Оценки коэффициента β_1 метода первых разностей и фиксированных эффектов совпадают.

Обсудим число степеней свободы для каждой из оценок.

Для метода первых разностей число наблюдений равно $N/2 = K$. Соответственно, число степеней свободы равно K минус число оцениваемых параметров:

$$Df = \frac{N}{2} - 1 = K - 1.$$

Для модели фиксированных эффектов число наблюдений равно N. Однако число параметров равно $K + 1 = N/2 + 1$. Число степеней свободы:

$$Df = N - (\frac{N}{2} + 1) = \frac{N}{2} - 1 = K - 1.$$

Итак, число степеней свободы для методов первых разностей и фиксированных эффектов совпадает. То есть в рассматриваемом примере

оценки, полученные путем перехода к разностям и фиксированных эффектов, полностью эквивалентны.

При применении метода первых разностей следует иметь в виду следующие обстоятельства:

- модель должна быть линейной;
- итоговое уравнение получается в виде разности;
- в случае ровно двух наблюдений в группе, оценка первых разностей полностью совпадает с оценкой уравнения с фиктивными переменными, соответствующими каждой паре.

В рассмотренном примере переход к разностям или использование фиксированных эффектов обосновывался тем, что были собраны данные о близнецах. Мы допустили, что для близнецов уровень способностей одинаков. А если бы пары формировались из братьев или сестер разного возраста, уже было бы невозможно утверждать, что ненаблюдаемые переменные принимают одинаковые значения внутри пары. Метод первых разностей или фиксированных эффектов позволяет учесть лишь ту неоднородность объектов, которая связана с ненаблюдаемыми переменными, которые постоянны внутри группы.

Когда панель строится по времени, часто удается обосновать, что некоторые скрытые параметры не изменяются от периода к периоду.

Пусть, например, цель исследования состоит в оценке влияния налоговых скидок на инвестиции (ITC) на стоимость акций.

Исходная модель формулируется следующим образом:

$$P_{ij} = \beta_0 + \beta_1 D_{ij}^{ITC} + \beta_2 \pi_j + \beta_3 X_{ij} + \varepsilon_{ij},$$

где P_{ij} - цена одной акции, j - номер фирмы, i – номер года (1 или 2);

D^{ITC}_{ij} – фиктивная переменная, равная 1, если фирма получала скидку в текущем году, 0 – в противном случае;

π_j - внутренние скрытые факторы прибыльности (качество менеджмента, и т.д.);

X_{ij} – прочие наблюдаемые характеристики фирмы или рынка (например, число конкурентов и т.п.).

Проблема состоит в том, что величины π_j не известны, и их нельзя оценить по данным отчетности. Кроме того, только те фирмы, которые имели прибыль, могут получить налоговые скидки. Поэтому

$$\mathrm{cov}(\pi_j, D_{ij}^{ITC}) \neq 0.$$

Если имеются панельные данные, включающие сведения о фирмах в год до получения налоговых скидок (год *t-1*) и последующий (год *t*), для построения оценки мы можем использовать метод первых разностей. При этом мы должны надеяться, что для каждой из фирм внутренние факторы доходности не изменились за этот период. Тогда

$$P_{jt} - P_{jt-1} = (\beta_0 - \beta_0) + \beta_1 (D_{it}^{ITC} - D_{it-1}^{ITC}) + \\ + \beta_2 (\pi_j - \pi_j) + \beta_3 (X_{it} - X_{it-1}) + (\varepsilon_{it} - \varepsilon_{it-1})$$

или

$$\Delta P_{it} = \beta_1 \Delta D_{it}^{ITC} + \beta_3 \Delta X_{it} + \Delta \varepsilon_{it}.$$

Если записать модель в этой форме, то можно получить несмещенную оценку β_1.

Эквивалентную оценку можно найти, используя модель фиксированных эффектов и добавив в модель для каждой фирмы (кроме последней или любой другой) фиктивную переменную.

Допустим мы хотим оценить склонность домохозяйств к сбережению. Запишем исходную модель:

$$S_{it} = \beta_0 + \beta_1 Y_{it} + \beta_2 F_{it} + \beta_3 X_{it} + \varepsilon_{it},$$

где S_{it} – сбережения i-го домохозяйства в год t;

Y_{it} – доход i-го домохозяйства в год t;

F_{it} – ненаблюдаемые характеристики i-го домохозяйства в год t (склонности, способность к предвидению и т.д.);

X_{it} – прочие характеристики i-го домохозяйства в год t (возраст главы семьи, количество детей и т.п.).

Величины F_{it} не поддаются непосредственному измерению и коррелированны с доходом:

$$\mathrm{cov}(F_i, Y_{it}) \neq 0.$$

Перейдя к первым разностям, получим:

$$S_{jt} - S_{jt-1} = (\beta_0 - \beta_0) + \beta_1(Y_{it} - Y_{it-1}) + \beta_2(F_j - F_j) + \\ + \beta_3(X_{it} - X_{it-1}) + (\varepsilon_{it} - \varepsilon_{it-1}).$$

или

$$\Delta S_{jt} = \beta_1 \Delta Y_{it} + \beta_3 \Delta X_{it} + \Delta\varepsilon_{it}.$$

Полученное уравнение не содержит ненаблюдаемых переменных. Используя метод наименьших квадратов, можно найти несмещенную оценку склонности к сбережению β_1. Альтернативный метод оценивания - введение в исходную модель фиктивных переменных – приведет к аналогичному результату.

В рассмотренных примерах все совокупности содержат некоторую внутреннюю неоднородность. Некоторые факторы скрыты, их не удается измерить и включить в модель. Панельные данные позволяют частично учесть эту неоднородность за счет того, что индивидуальные эффекты отражают влияние всех (наблюдаемых или ненаблюдаемых) переменных, которые принимают разные значения для разных объектов, но не меняются во времени. Аналогично, если добавить в модель фиктивные переменные для каждого момента времени, то коэффициенты при них вберут в себя влияния всех наблюдаемых или ненаблюдаемых переменных, которые зависят только от времени, но одинаковы для всех единиц совокупности.

Если ненаблюдаемые переменные коррелированны с регрессорами, то оценки коэффициентов сокращенной модели будут смещены (из-за эффекта пропущенных переменных). Использование панельных данных позволяет устранить это смещение. То есть оценки параметров, рассчитанные по панельным данным более робастны – устойчивы по отношению к неполной спецификации модели. Соответствующие методы получили названия метода первых разностей (*FD*), вторых разностей (*DD*), модели с фиксированными эффектами (*fixed effects*).

Если допустить, что пропущенные переменные не коррелированны с остальными регрессорами, тогда их влияние можно учесть иначе -

рассматривать как компоненты ошибок наблюдения. Тогда для панельных данных используют ***модели со случайными эффектами (random effects models)***. Основанием такого допущения могут быть положения экономической теории или особенности организации выборки. Если пропущенные переменные являются одной из составляющих ошибок, получим:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \varepsilon_{it},$$

$$\varepsilon_{it} = \nu_i + \omega_{it},$$

$$\nu_i \sim IID(0, \sigma_\nu^2),\ \omega_{it} \sim IID(0, \sigma_\omega^2),$$

где ε_{it} – суммарная ошибка;

ν_i – ошибка, характерная для *i*-го объекта и не зависящая от времени;

ω_{it} – случайная ошибка регистрации.

Это модель линейной регрессии при гетероскедастичности ошибок. Дисперсия ошибок зависит от номера объекта. Поэтому для оценивания следует использовать обобщенный метод наименьших квадратов.

Такая модель обладает рядом преимуществ. Во-первых, она содержит меньше параметров, чем модель с фиксированными эффектами. Во-вторых, если предположение о некоррелированности индивидуальных эффектов и регрессоров выполняется, оценки модели со случайными эффектами более эффективны. Но, если предположение не верно, то оценки будут смещены.

Поясним, как возникает смещение для случая единственной независимой переменной *X*. Рассмотрим модель:

$$Y_{it} = \mu_i + \beta_1 X_{it} + \varepsilon_{it}.$$

Допустим, что все индивидуальные эффекты μ_i и регрессионный коэффициент β_1 - известны, значения переменной *X* сильно различаются для разных объектов, но изменения от раунда к раунду не велики. Посмотрим на диаграмму рассеивания (*X, Y*) для случаев:

1) когда индивидуальные эффекты коррелированны с независимой переменной *X* и

2) когда индивидуальные эффекты не коррелированны с независимой переменной X.

Параметры μ_i задают семейство N параллельных прямых. На рис.4 представлен случай, когда индивидуальные эффекты коррелированны с независимой переменной X. Упорядочим объекты по величине среднего значения независимой переменной. Тогда для первого объекта значения X – самые меньшие. Так как индивидуальный эффект коррелирован с $\bar{X}_1$, то μ_1 – также будет мало. На рисунке возможные реализации значений для первого объекта группируются вокруг самой нижней прямой. Для наглядности они обведены самым нижним овалом. Для следующего объекта значения независимой переменной и индивидуальный эффект – чуть больше и т.д. Большему среднему значению X для i-го объекта будет соответствовать большее значение индивидуального эффекта μ_i.

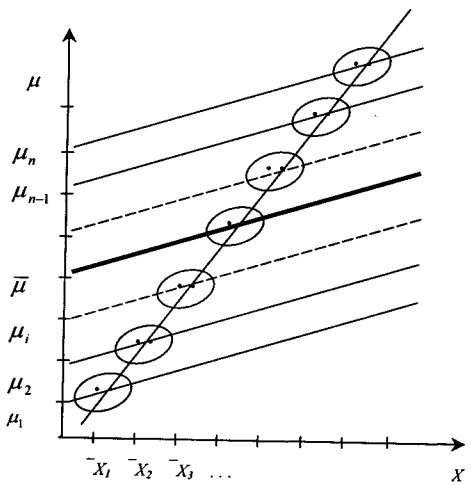

Рис. 4. Индивидуальные эффекты коррелированны с независимой переменной

Если по этим точкам мы построим уравнение регрессии, неважно обычным методом наименьших квадратов или обобщенным методом наименьших квадратов (модель случайных эффектов), оценка

коэффициента наклона будет завышена, то есть прямая пройдет под большим углом, чем истинная зависимость. Но модель с фиксированными эффектами позволяет получить несмещенную оценку коэффициента β_1.

На рис. 5 представлен случай, когда величина индивидуального эффекта не зависит значений X.

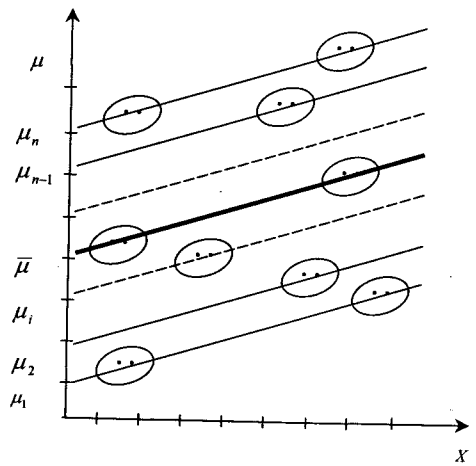

Рис. 5. Отсутствие корреляции между индивидуальными эффектами и регрессорами

В этом случае можно пользоваться всеми тремя моделями – обычной регрессии, фиксированных эффектов, случайных эффектов, но наиболее эффективные оценки даст модель случайных эффектов.

Более подробно модели с фиксированными и случайными эффектами мы рассмотрим в следующем разделе.

2. МОДЕЛЬ С ФИКСИРОВАННЫМИ ЭФФЕКТАМИ

2.1. Оценивание коэффициентов модели

Рассмотрим модель линейной регрессии для панельных данных включающую индивидуальные уровни для каждого объекта:

$$y_{it} = \alpha_i + \beta' \mathbf{x}_{it} + \varepsilon_{it},\ i=1,2,\ldots,N,\ t=1,2,\ldots,T. \qquad (1)$$

$$\varepsilon_{it} \sim IID(0,\sigma^2),\ \mathrm{cov}(\varepsilon_{it},\varepsilon_{js}) = 0,\ i \neq j, t \neq s.$$

Для каждого объекта $i=1,\ldots,N$ индивидуальный эффект α_i остается постоянным в течение всех периодов $t=1,\ldots T$. Вектор регрессоров $\mathbf{x}_{it}=(\mathbf{x}^1_{it}, \mathbf{x}^2_{it},\ldots, \mathbf{x}^k_{it})'$ не включает свободного члена. Таким образом, в (1) α_i, $i=1,\ldots,N$ - неизвестные параметры, которые необходимо оценить.

Уравнение (1) можно записать в векторной и матричной форме. Обозначим y_i – вектор размерности T значений независимых переменных для i-го объекта; X_i – матрица значений регрессоров для i-го объекта размерности $T \times K$, и пусть ε_i - вектор ошибок размерности $T \times 1$.

Тогда (1) можно переписать в виде:

$$y_i = \mathbf{i}\alpha_i + X_i\beta + \varepsilon_i,\ i=1,\ldots N.$$

где $\mathbf{i}$ – вектор, состоящий из единиц, размерности T.

Объединяя уравнения в единую систему, получим:

$$\begin{bmatrix} y_1 \\ y_2 \\ M \\ y_N \end{bmatrix} = \begin{bmatrix} \mathbf{i} & 0 & \Lambda & 0 \\ 0 & \mathbf{i} & \Lambda & 0 \\ & & M & \\ 0 & 0 & \Lambda & \mathbf{i} \end{bmatrix} \cdot \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ M \\ \alpha_N \end{bmatrix} + \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ M \\ \mathbf{X}_N \end{bmatrix} \beta = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ M \\ \varepsilon_N \end{bmatrix}$$

или

$$\mathbf{y} = [\mathbf{d}_1 \quad \mathbf{d}_2 \quad \ldots \quad \mathbf{d}_N \ \mathbf{X}] \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \varepsilon, \qquad (2)$$

где $\mathbf{d}_i$ - фиктивная переменная, соответствующая i-му объекту.

Если обозначить $\mathbf{D} = [\mathbf{d}_1\ \mathbf{d}_2 \ldots\ \mathbf{d}_N]$ - матрица размером $NT \times N$,

получим матричную запись:

$$\mathbf{y} = \mathbf{D}\alpha + X\beta + \varepsilon. \qquad (3)$$

Матрица X содержит K столбцов, матрица $\mathbf{D}$ - N столбцов, всего модель содержит $K+N$ оцениваемых параметров. Если число объектов N невелико, то оценки параметров $\mathbf{a} = \hat{\alpha}$ и $\mathbf{b} = \hat{\beta}$ можно получить с помощью стандартных формул регрессионного анализа.

Система нормальных уравнений имеет вид:

$$\begin{bmatrix} \mathbf{D'D} & \mathbf{D'X} \\ \mathbf{D'X} & \mathbf{X'X} \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{D'y} \\ \mathbf{X'y} \end{bmatrix}.$$

Для того, чтобы было возможно найти решение системы уравнений, необходимо, чтобы матрица системы имела полный ранг. Это означает, что регрессоры не должны быть коллинеарны с фиктивными переменными. В частности, матрица $\mathbf{X}$ не должна включать признаков, которые не изменяются во времени. Например, таких как год рождения или пол респондента. Если это требование выполнено, то оценки можно найти по формуле:

$$\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{D'D} & \mathbf{D'X} \\ \mathbf{D'X} & \mathbf{X'X} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{D'y} \\ \mathbf{X'y} \end{bmatrix}$$

Но, если число единиц наблюдения составляет несколько сотен или тысяч, вычисление обратной матрицы потребует слишком больших затрат времени и объемов оперативной памяти. Чтобы избежать этого учитывают, что матрица $\mathbf{D}$ состоит из значений N фиктивных переменных, а регрессионная модель является ***моделью регрессии с фиктивными переменными (LSDV)***.

Для моделей с фиктивными переменными известно, что МНК-оценку $\mathbf{b}$ для вектора параметров можно найти из линейной регрессии для преобразованных переменных $\mathbf{X^*} = \mathbf{M_dX}$ и $\mathbf{y^*} = \mathbf{M_dX}$, где

$$\mathbf{M_d} = \mathbf{I} - \mathbf{D}(\mathbf{D'D})^{-1}\mathbf{D'}.$$

Оценка **b** вычисляется по формуле:

$$\mathbf{b}= [\mathbf{X}'\mathbf{M}_d\mathbf{X}]^{-1}[\mathbf{X}'\mathbf{M}_d\mathbf{y}], \quad (4)$$

Рассмотрим, что означает предварительное преобразование переменных. Так как столбцы матрицы **D** ортогональны, то матрица $\mathbf{M}_d$ является блочно-диагональной:

$$\mathbf{M_d} = \begin{bmatrix} \mathbf{M}^0 & 0 & 0 & \Lambda & 0 \\ 0 & \mathbf{M}^0 & 0 & \Lambda & 0 \\ & & \mathbf{M} & & \\ 0 & 0 & 0 & \Lambda & \mathbf{M}^0 \end{bmatrix},$$

где $\mathbf{M}^0$ – матрица вида:

$$\mathbf{M}^0 = \mathbf{I_T} - \frac{1}{T}\mathbf{i}\mathbf{i}',$$

где $\mathbf{I_T}$ – единичная матрица размером $T \times T$;

i – вектор, размерности T, все элементы которого равны единице.

$$\mathbf{M}^0 = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix} - \begin{bmatrix} 1/T & 1/T & \dots & 1/T \\ 1/T & 1/T & \dots & 1/T \\ \dots & \dots & \dots & \dots \\ 1/T & 1/T & \dots & 1/T \end{bmatrix}$$

Для любого вектора $\mathbf{z}_i$ размерности $T \times 1$ умножение на $\mathbf{M}^0$, означает вычитание из каждого его элемента среднего значения: $\mathbf{M}^0\mathbf{z} = \mathbf{z} - \frac{1}{T}\sum_{t=1}^{T} z_t\mathbf{i} = \mathbf{z} - \bar{\mathbf{z}}\mathbf{i}$. То есть преобразование $\mathbf{M_d y}$ и $\mathbf{M_d X}$ означает вычитание из каждого элемента вектора **y** и матрицы **X** среднего значения: $[y_{it} - \bar{y}_i]$ и $[x_{it} - \bar{x}_i]$, где $\bar{y}_i$ - среднее значение зависимой переменной для i-го объекта, $\bar{x}_i$ - вектор средних значений независимых переменных для i-го объекта.

После того как по формуле (4) найдены оценки **b** легко получить и оценки индивидуальных эффектов **a** из уравнения:

$$\mathbf{D}'\mathbf{D}\mathbf{a} + \mathbf{D}'\mathbf{X}\mathbf{b} = \mathbf{D}'\mathbf{y}$$

или

$$\boldsymbol{a} = [\mathbf{D}'\mathbf{D}]^{-1}\mathbf{D}'(\mathbf{y}-\mathbf{X}\mathbf{b}). \quad (5)$$

Формула (5) означает, что для каждого объекта индивидуальный эффект можно найти по формуле:

$$a_i = \bar{\mathbf{y}}_i - \mathbf{b}'\bar{\mathbf{x}}_i. \quad (6)$$

То есть индивидуальный эффект a_i равен среднему остатку в i-ой группе.

Оценка ковариационной матрицы вектора **b** может получена по обычной формуле:

$$\text{Est. Var}[\mathbf{b}] = s^2[\mathbf{X}'\mathbf{M}_d\mathbf{X}]^{-1}, \quad (7)$$

где s^2 - несмещенная оценка остаточной дисперсии:

$$s^2 = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - a_i - \mathbf{x}'_{it}\mathbf{b})^2}{NT - N - K}. \quad (8)$$

Если учесть, что для каждого остатка выполняется равенство:

$$\begin{aligned} e_{it} &= y_{it} - a_i - \mathbf{x}'_{it}\mathbf{b} = \\ &= y_{it} - (\bar{y}_i - \bar{\mathbf{x}}'_i\mathbf{b}) - \mathbf{x}'_{it}\mathbf{b} = \\ &= (y_{it} - \bar{y}_i) - (\mathbf{x}_{it} - \bar{\mathbf{x}}\,\mathbf{i})'\mathbf{b}, \end{aligned}$$

очевидно, что числитель в выражении (8) равен сумме квадратов остатков регрессии (4).

Таким образом, вычисления для модели с фиксированными эффектами можно провести с помощью стандартных пакетов статистического анализа. Для этого необходимо найти для всех входящих в модель переменных средние значения для каждого объекта за весь период наблюдения. Вычесть групповые средние из исходных данных. К преобразованным данным можно применять стандартные методы регрессионного анализа. Коэффициенты совпадут с моделью фиксированных эффектов. Но, оценку остаточной дисперсии, стандартные ошибки коэффициентов, F и t-статистик необходимо

скорректировать. Различия вызваны тем, что в качестве знаменателя при вычислении остаточной дисперсии s^2 будет использоваться $(NT - K)$ вместо $(NT - N - K)$. Этот факт необходимо учитывать и вносить соответствующие поправки.

Оценки выборочных дисперсий индивидуальных эффектов можно получить по формуле:

$$Var[a_i] = \frac{\sigma^2}{T} + \bar{\mathbf{x}}'_i Var[\mathbf{b}] \mathbf{x}'_i,$$

Пример 1. Пусть матрица данных содержит сведения о пяти объектах за три последовательные периода времени.

Таблица 2

Исходные данные					Средние за три периода		
i	t	$x1_{it}$	$x2_{it}$	y_{it}	$\bar{x}1_i$	$\bar{x}2_i$	$\bar{y}_i$
1	1	3	10	3,3	3	9	2,7
1	2	4	10	1,9			
1	3	2	7	2,9			
2	1	5	7	3,3	5	7	3,5
2	2	4	6	4,3			
2	3	6	8	2,9			
3	1	0	11	12,9	0	12	13
3	2	0	12	12,8			
3	3	0	13	13,3			
4	1	1	12	14,3	2	14	14,4
4	2	4	13	12			
4	3	1	17	16,9			
5	1	4	12	14,4	6	15	14,2
5	2	5	14	14,8			
5	3	9	19	13,4			

Модель имеет вид:

$$y_{it} = \alpha_i + \beta_1 x1_{it} + \beta_2 x2_{it} + \varepsilon_{it}.$$

Стандартная регрессионная модель, включающая единственный свободный член, то есть $a_i = a$, приводит к уравнению:

$$\hat{y}_{it} = -2{,}61 - 0{,}77 x1 + 1{,}28 x2, \; RSS = 104.415, \; s^2 = 8{,}7013.$$

Рассмотрим модель с фиксированными эффектами. В наших обозначениях каждому объекту соответствуют следующие вектора и матрицы:

$$\mathbf{y}_1 = \begin{pmatrix} 3{,}3 \\ 1{,}9 \\ 2{,}9 \end{pmatrix}, \mathbf{X}_1 = \begin{pmatrix} 3 & 10 \\ 4 & 10 \\ 2 & 7 \end{pmatrix}, \mathbf{y}_2 = \begin{pmatrix} 3{,}3 \\ 4{,}3 \\ 2{,}9 \end{pmatrix}, \mathbf{X}_2 = \begin{pmatrix} 5 & 7 \\ 4 & 6 \\ 6 & 8 \end{pmatrix}, \mathbf{y}_3 = \begin{pmatrix} 12{,}9 \\ 12{,}8 \\ 13{,}3 \end{pmatrix}, \mathbf{X}_3 = \begin{pmatrix} 0 & 11 \\ 0 & 12 \\ 0 & 13 \end{pmatrix},$$

$$\mathbf{y}_4 = \begin{pmatrix} 14{,}3 \\ 12{,}0 \\ 16{,}9 \end{pmatrix}, \mathbf{X}_4 = \begin{pmatrix} 1 & 12 \\ 4 & 13 \\ 1 & 17 \end{pmatrix}, \mathbf{y}_5 = \begin{pmatrix} 14{,}4 \\ 14{,}8 \\ 13{,}4 \end{pmatrix}, \mathbf{X}_5 = \begin{pmatrix} 4 & 12 \\ 5 & 14 \\ 9 & 19 \end{pmatrix}, \mathbf{i} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Матрицы **D и X** равны:

$$\mathbf{D} = \begin{pmatrix} 1&0&0&0&0 \\ 1&0&0&0&0 \\ 1&0&0&0&0 \\ 0&1&0&0&0 \\ 0&1&0&0&0 \\ 0&1&0&0&0 \\ 0&0&1&0&0 \\ 0&0&1&0&0 \\ 0&0&1&0&0 \\ 0&0&0&1&0 \\ 0&0&0&1&0 \\ 0&0&0&1&0 \\ 0&0&0&0&1 \\ 0&0&0&0&1 \\ 0&0&0&0&1 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 3&10 \\ 4&10 \\ 2&7 \\ 5&7 \\ 4&6 \\ 6&8 \\ 0&11 \\ 0&12 \\ 0&13 \\ 1&12 \\ 4&13 \\ 1&17 \\ 4&12 \\ 9&14 \\ 5&19 \end{pmatrix}.$$

Система нормальных уравнений имеет вид:

$$\begin{bmatrix} 3&0&0&0&0&9&26 \\ 0&3&0&0&0&15&21 \\ 0&0&3&0&0&0&36 \\ 0&0&0&3&0&6&43 \\ 0&0&0&0&3&18&44 \\ 9&15&0&6&18&246&561 \\ 27&21&36&42&45&561&2135 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 8{,}1 \\ 10{,}5 \\ 39 \\ 43{,}2 \\ 42{,}7 \\ 405{,}8 \\ 1862{,}3 \end{bmatrix}.$$

Размерность системы уравнений составляет $N+K=5+2=7$. Решая систему, найдем оценки параметров:

$$\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 1{,}21 \\ 4{,}92 \\ 7{,}13 \\ 9{,}49 \\ 12{,}68 \\ -0{,}97 \\ 0{,}49 \end{bmatrix}.$$

Однако прямой способ расчета можно применять лишь при небольшом числе объектов.

Оценки b_1 и b_2 можно получить, используя формулу (4). Для этого найдем средние значения признаков для каждого объекта за три периода:

$$\bar{\mathbf{x}}\mathbf{1}_{i\bullet} = \begin{bmatrix} 3 \\ 5 \\ 0 \\ 2 \\ 6 \end{bmatrix}, \ \bar{\mathbf{x}}\mathbf{2}_{i\bullet} = \begin{bmatrix} 9 \\ 7 \\ 12 \\ 14 \\ 15 \end{bmatrix}, \bar{\mathbf{y}}_{i\bullet} = \begin{bmatrix} 2{,}7 \\ 3{,}5 \\ 13 \\ 14{,}4 \\ 14{,}2 \end{bmatrix}.$$

Отклонения от средних значений за три периода для каждого объекта приведены в таблице:

i	t	$x1_{it} - \bar{x}1_i$	$x2_{it} - \bar{x}2_i$	$y_{it} - \bar{y}_i$
1	1	0	1	0,6
1	2	1	1	-0,8
1	3	-1	-2	0,2
2	1	0	0	-0,2
2	2	-1	-1	0,8
2	3	1	1	-0,6
3	1	0	-1	-0,1
3	2	0	0	-0,2
3	3	0	1	0,3
4	1	-1	-2	-0,1
4	2	2	-1	-2,4
4	3	-1	3	2,5
5	1	-2	-3	0,2
5	2	-1	-1	0,6
5	3	3	4	-0,8

Получим оценки коэффициентов, используя стандартные формулы метода наименьших квадратов:

$$\mathbf{b} = \begin{bmatrix} 24 & 21 \\ 21 & 50 \end{bmatrix}^{-1} \begin{bmatrix} -13 \\ 4 \end{bmatrix} = \begin{bmatrix} -0{,}97 \\ 0{,}49 \end{bmatrix}.$$

Индивидуальные эффекты найдем по формуле:

$$a_i = \bar{y}_i - \mathbf{b}'\bar{\mathbf{x}}_i.$$

$$a_1 = 2{,}7-(-0{,}97\times 3+0{,}49\times 9) = 1{,}2,$$

$$a_2 = 3{,}5-(-0{,}97\times 5+0{,}49\times 7) = 4{,}92,$$

$$a_3 = 13-(-0{,}97\times 0+0{,}49\times 12) = 7{,}12,$$

$$a_4 = 14{,}4-(-0{,}97\times 2+0{,}49\times 14) = 9{,}48,$$

$$a_5 = 14{,}(-0{,}97\times 6+0{,}49\times 15) = 12{,}67.$$

Сумма квадратов остатков равна:

$$RSS = 0{,}665982.$$

Оценка дисперсии ошибок:

$$s^2 = 0{,}6659/(15-7) = 0{,}0832.$$

Оба способа нахождения коэффициентов полностью эквивалентны и приводят к одинаковым результатам.

Рассчитаем стандартные ошибки коэффициентов регрессии. Воспользуемся формулой $s_{b_j} = \sqrt{s^2[(X'X)^{-1}]_{jj}}$. Элементы обратной матрицы равны:

$$\begin{bmatrix} 24 & 21 \\ 21 & 50 \end{bmatrix}^{-1} = \begin{bmatrix} 0.0659 & -0.0277 \\ -0.0277 & 0.0316 \end{bmatrix}$$

Следовательно,

$$s_{b_1} = \sqrt{0.0832 \cdot 0.0659} = 0.0741,$$

$$s_{b_2} = \sqrt{0.0832 \cdot 0.0316} = 0.0513.$$

Построим с надежностью $\gamma=0{,}95$ интервальные оценки параметров:

$$\beta_j \in \{b_j \pm t_\gamma s_{b_j}\},$$

$$\beta_1 \in [-0.9698 \pm 2.306 \times 0.0741],$$

$$\beta_2 \in [-.48932 \pm 2.306 \times 0.0513],$$

где t_γ= 2,306 находим по таблице t-распределения при a = 1-γ= 0.95 и $\nu = N^*T - N - K = 15 - 7 = 8$.

Следовательно,

$$[-1.1406 \le \beta_1 \le -0.7991],$$

$$[0.3710 \le \beta_2 \le 0.6076].$$

Значимость коэффициентов β_1 и β_2 можно проверить, вычислив t-статистики.

2.2. Проверка гипотезы о значимости групповых эффектов

Для того, чтобы доказать, что введение в модель фиксированных эффектов оправдано, необходимо проверить гипотезу об их значимости. Если они равны между собой, то модели с фиксированными эффектами следует предпочесть обычную регрессию. Для проверки используется F-статистика:

$$F(N-1, NT-N-K) = \frac{(RSS^r - RSS^{FE})/(N-1)}{RSS^{FE}/(NT-N-K)}, \qquad (9)$$

которая имеет $F(N\text{-}1, NT\text{-}N\text{-}K)$ распределение с N-1 и NT-N-K степенями свободы; RSS^{FE} - сумма квадратов остатков модели с фиксированными эффектами, RSS^r - модели, включающей лишь единственный свободный член.

Если модель переписать в форме, включающей свободный член α_0 и (N-1) фиктивную переменную $\alpha_2, \ldots, \alpha_N$, то для оценки индивидуального эффекта вместо α_i необходимо использовать разность α_i-α_0. Все основные результаты в этом случае не изменятся. F-статистика для проверки гипотезы о равенстве (N-1) коэффициентов при фиктивных переменных

$\alpha_2, \ldots, \alpha_N$ нулю вычисляется по той же самой формуле. Отличия касаются лишь интерпретации коэффициентов α_i.

Для данных примера 1 RSS^r= 104,415, RSS^{FE} = 0,6659, вычисленное и критическое значение F-статистики составляют:

$$F_{набл}(4,8) = [(104{,}415 - 0{,}6659) / (5\text{-}1)] / [0{,}6659 / (15\text{-}7)] = 311{,}57,$$

$$F_{крит}(4,8,\alpha=0{,}01) = 7{,}01.$$

Наблюдаемое значение попадает в критическую область, гипотеза о равенстве всех индивидуальных эффектов H_0: $a_1 = a_2 = a_3 = a_4 = a_5$ отвергается.

С помощью F-критерия можно проверить гипотезу H_0: β_1=β_2= 0:

$$F(K, NT-N-K) = \frac{(TSS^{FE} - RSS^{FE})/K}{RSS^{FE}/(NT-N-K)},$$

где $TSS^{FE} = \sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \bar{y}_i)^2$. В рассматриваемом примере TSS^{FE} = 15,28, $F_{набл}(2,8)$ = 87,77, $F_{крит}(2,8,\alpha$=0,01) = 8,6. Следовательно, коэффициенты значимо отличаются от нуля.

2.3. Оценки "within" и "between"

Как мы видели оценки модели с фиксированными эффектами можно получить, переходя к отклонениям от групповых средних. Рассмотрим как взаимосвязаны оценки трех различных регрессий.

Во-первых, с единственным свободным членом:

$$y_{it} = \alpha + \beta' \mathbf{x}_{it} + \varepsilon_{it}. \qquad (10a)$$

Во-вторых, построенной по отклонениям от групповых средних:

$$y_{it} - \bar{y}_i = \beta'(\mathbf{x}_{it} - \bar{\mathbf{x}}_i) + \varepsilon_{it} - \bar{\varepsilon}_i. \qquad (10b)$$

В-третьих, по групповым средним:

$$\bar{y}_i = \alpha + \beta' \bar{\mathbf{x}}_i + \bar{\varepsilon}_i. \qquad (10c)$$

Каждая из этих возможностей играет важную роль при анализе панельных данных и используется напрямую либо на промежуточных

этапах. Проанализируем, как взаимосвязаны МНК-оценки коэффициентов этих трех уравнений. Для этого рассмотрим, какие матрицы сумм квадратов и перекрестных произведений используются в каждом случае.

В (10a) перекрестные произведения находятся от общих средних, $\bar{\bar{y}}$ и $\bar{\bar{x}}$:

$$S_{xx}^{t} = \sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{\bar{x}})(x_{it} - \bar{\bar{x}})'$$

и

$$S_{xy}^{t} = \sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{\bar{x}})(y_{it} - \bar{\bar{y}})'.$$

В случае модели (10b) данные представлены в виде отклонений от групповых средних, поэтому выборочные средние $(y_{it} - \bar{y}_i)$ и $(x_{it} - \bar{x}_i)$ равны нулю. Матрицы перекрестных произведений рассчитываются по отклонениям групповых средних и отражают внутригрупповые суммы квадратов:

$$S_{xx}^{w} = \sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)'$$

и

$$S_{xy}^{w} = \sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i).$$

Наконец, в (10c) среднее групповых средних является общим средним. Матрицы моментов равны:

$$S_{xx}^{b} = \sum_{i=1}^{N}T(\bar{x}_i - \bar{\bar{x}})(\bar{x}_i - \bar{\bar{x}})'$$

и

$$S_{xy}^{b} = \sum_{i=1}^{N}T(\bar{x}_i - \bar{\bar{x}})(\bar{y}_i - \bar{\bar{y}}).$$

Легко проверить, что выполняются равенства:

$$S_{xx}^{t} = S_{xx}^{w} + S_{xx}^{b}$$

и

$$S_{xy}^{t} = S_{xy}^{w} + S_{xy}^{b}.$$

Найдем оценки неизвестных параметров β каждым из трех способов.

Оценки стандартной регрессии:

$$\boldsymbol{b}^{t} = [S_{xx}^{t}]^{-1}S_{xy}^{t} = [S_{xx}^{w} + S_{xx}^{b}]^{-1}[S_{xy}^{w} + S_{xy}^{b}]. \qquad (11)$$

Основываясь отклонениях от групповых средних, получим оценки:

$$\boldsymbol{b}^{w} = [S_{xx}^{w}]^{-1}S_{xy}^{w}. \qquad (12)$$

Полученные данным способом оценки иногда называют ***within оценками или внутригрупповыми (within-groups estimator).*** Как уже было показано, эта формула соответствует модели с фиксированными эффектами.

Оценки наименьших квадратов для уравнения (l0c), рассчитываемые по N групповым средним:

$$\boldsymbol{b}^{b} = [S_{xx}^{b}]^{-1}S_{xy}^{b}. \qquad (13)$$

называют ***between оценками или межгрупповыми (between-groups estimator).***

Из выражений (12) и (13) следует, что:

$$S_{xy}^{w} = S_{xx}^{w}\boldsymbol{b}^{w}$$

и

$$S_{xy}^{b} = S_{xx}^{b}\boldsymbol{b}^{b}.$$

Включая их в (11), мы видим, что МНК оценка есть взвешенная сумма внутригрупповых и межгрупповых оценок:

$$b^t = F^w b^w + F^b b^b, \qquad (14)$$

где

$$F^w = [S^w_{xx} + S^b_{xx}]^{-1} S^w_{xx} = \mathbf{I} - F^b,$$

$$F^b = [S^w_{xx} + S^b_{xx}]^{-1} S^b_{xx}.$$

Пример 1 (продолжение).

Метод заводо-лет приводит к следующим матрицам сумм квадратов и перекрестных произведений:

$$S^t_{xx} = \begin{bmatrix} 92,4 & 13,8 \\ 13,8 & 185,6 \end{bmatrix}, \quad S^t_{xy} = \begin{bmatrix} -53,08 \\ 227,54 \end{bmatrix}, \quad b^t = \begin{bmatrix} -0,77 \\ 1,28 \end{bmatrix}.$$

Оценка по отклонениям от групповых средних:

$$S^w_{xx} = \begin{bmatrix} 24 & 21 \\ 21 & 50 \end{bmatrix}, \quad S^w_{xy} = \begin{bmatrix} -13 \\ 4 \end{bmatrix}, \quad b^w = \begin{bmatrix} -0,97 \\ 0,49 \end{bmatrix}.$$

Регрессия по средним внутри пяти групп:

$$S^b_{xx} = \begin{bmatrix} 68,4 & -7,2 \\ -7,2 & 135,6 \end{bmatrix}, \quad S^b_{xy} = \begin{bmatrix} -40,08 \\ 223,44 \end{bmatrix}, \quad b^b = \begin{bmatrix} -0,41 \\ 1,63 \end{bmatrix}.$$

Эти три оценки связаны следующим образом:

$$b^t = F^w b^w + F^b b^b = \begin{bmatrix} -0,77 \\ 1,28 \end{bmatrix} = \begin{bmatrix} 0,25 & 0,19 \\ 0,09 & 0,25 \end{bmatrix} \begin{bmatrix} -0,97 \\ 0,49 \end{bmatrix} + \begin{bmatrix} 0,75 & -0,19 \\ -0,09 & 0,74 \end{bmatrix} \begin{bmatrix} -0,41 \\ 1,63 \end{bmatrix} = \begin{bmatrix} -0,77 \\ 1,28 \end{bmatrix}.$$

2.4. Двунаправленная модель с фиксированными эффектами

Фиктивные переменные могут использоваться и для учета временного фактора. Это необходимо, если средний уровень явления существенно изменяется во времени. Тогда модель может быть расширена следующим образом:

$$y_{it} = \alpha_i + \gamma_t + \beta' \mathbf{x}_{it} + \varepsilon_{it}, \qquad (15)$$

где γ_t – эффект специфический для каждого периода. Эта модель получена из предыдущей включением дополнительных T - 1 фиктивных переменных. Из-за строгой коллинеарности нельзя включить все T эффектов для каждого периодов времени. На практике обычно исключают эффект для первого γ_1 или последнего периода γ_T).

Модель также можно переписать в симметричной форме:

$$y_{it} = \mu + \alpha_i + \gamma_t + \beta' \mathbf{x}_{it} + \varepsilon_{it}, \qquad (15')$$

с ограничениями

$$\sum_i \alpha_i = \sum_t \gamma_t = 0.$$

Будем считать, что для выполнения ограничений мы не включаем в модель эффекты α_N и γ_T. То есть, полагаем:

$$a_N = -\sum_{i=1}^{N-1} a_i, \quad \gamma_T = -\sum_{i=1}^{T-1} \gamma_i.$$

Обозначив $\mathbf{D_N}$, $\mathbf{D_T}$ – множества из N-1и T-1 фиктивных переменных получим матричную запись:

$$\mathbf{y} = \mathbf{i}\mu + \mathbf{D_N}\alpha + \mathbf{D_T}\gamma + \mathrm{X}\beta + \varepsilon.$$

При этом матрица $\mathbf{D} = [\mathbf{i} \ \mathbf{D_N} \ \mathbf{D_T} \ \mathbf{X}]$ должна иметь ранг $N+T+K-1$. Это накладывает ограничения на состав регрессоров. В их число не должны входить не только постоянные во времени переменные, такие как пол или год рождения, но и зависящие только от времени и одинаковые для всех объектов, такие как уровень цен или темп инфляции.

Использование прямых формул предполагает вычисление обратной матрицы: $(\mathbf{D}'\mathbf{D})^{-1}$, что при большом числе объектов может потребовать слишком много времени. Однако вместо этого можно использовать свойства модели с фиктивными переменными и перейти к отклонениям от групповых средних. Для этого вычислим:

$$y_{*it} = y_{it} - \bar{y}_i - \bar{y}_t + \bar{\bar{y}}, \qquad (16)$$

$$x_{*it} = x_{it} - \bar{x}_i - \bar{x}_t + \bar{\bar{x}},$$

где

$$\bar{y}_t = \frac{1}{N}\sum_{i=1}^{N} y_{it},$$

$$\bar{\bar{y}} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it},$$

$$\bar{x}_t = \frac{1}{N}\sum_{i=1}^{N} x_{it},$$

$$\bar{\bar{x}} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}.$$

Найдем коэффициенты регрессии y_{*it} на x_{*it}. В результате получим оценку вектора **b**. Оценки коэффициентов при фиктивных переменных рассчитаем по формулам:

$$m = \hat{\mu} = \bar{\bar{y}} - \mathbf{b}'\bar{\bar{\mathrm{x}}},$$

$$a_i = (\bar{y}_i - \bar{\bar{y}}) - \mathbf{b}'(\bar{\mathrm{x}}_i - \bar{\bar{\mathrm{x}}}), \qquad (17)$$

$$c_t = \hat{\gamma}_t = (\bar{y}_t - \bar{\bar{y}}) - \mathbf{b}'(\bar{\mathrm{x}}_t - \bar{\bar{\mathrm{x}}}).$$

Для вычисления оценки ковариационной матрицы для вектора **b** используют суммы квадратов и перекрестных произведений $\mathbf{x}_{*it}$. Оценка дисперсии ошибок s^2 находится по формуле:

$$s^2 = \frac{RSS}{NT-(N-1)-(T-1)-1} = \frac{\mathbf{e}'\mathbf{e}}{NT-N-T+1}.$$

Если число периодов наблюдения невелико, то нет необходимости переходить к отклонениям от средних. Проще ввести в модель *T-1* фиктивную переменную для каждого из периодов за исключением первого или последнего.

Пример 1 (продолжение).

Средние значения признаков за периоды t=1,2,3 равны:

$$\bar{\mathbf{x}}\mathbf{1}_{*_t} = \begin{bmatrix} 2{,}6 \\ 3{,}4 \\ 3{,}6 \end{bmatrix}, \quad \bar{\mathbf{x}}2_{*_t} = \begin{bmatrix} 10{,}4 \\ 11 \\ 12{,}8 \end{bmatrix}, \quad \bar{\mathbf{y}}_{*_t} = \begin{bmatrix} 9{,}64 \\ 9{,}16 \\ 9{,}88 \end{bmatrix}.$$

Отклонения от средних приведены в таблице.

i	t	$x1_{*it} = x1_{it} - \bar{x}1_i - \bar{x}1_t + \bar{x}1$	$x2_{it} = x2_{*it} - \bar{x}2_i - \bar{x}2_t + \bar{x}2$	$y_{it} = y_{*it} - \bar{y}_i - \bar{y}_t + \bar{y}$
1	1	0,6	2	0,52
1	2	0,8	1,4	-0,4
1	3	-1,4	-3,4	-0,12
2	1	0,6	1	-0,28
2	2	-1,2	-0,6	1,2
2	3	0,6	-0,4	-0,92
3	1	0,6	0	-0,18
3	2	-0,2	0,4	0,2
3	3	-0,4	-0,4	-0,02
4	1	-0,4	-1	-0,18
4	2	1,8	-0,6	-2
4	3	-1,4	1,6	2,18
5	1	-1,4	-2	0,12
5	2	-1,2	-0,6	1
5	3	2,6	2,6	-1,12

Оценки коэффициентов:

$$\mathbf{b} = \begin{bmatrix} 21{,}2 & 15{,}6 \\ 15{,}6 & 34{,}4 \end{bmatrix}^{-1} \begin{bmatrix} -13 \\ 1{,}46 \end{bmatrix} = \begin{bmatrix} -0{,}97 \\ 0{,}48 \end{bmatrix}.$$

Свободный член и индивидуальные эффекты равны:

$$m = \hat{\mu} = \bar{\bar{y}} - \mathbf{b}'\bar{\bar{\mathrm{x}}} = 9{,}56 - (-0{,}97\times 3{,}2 + 0{,}48\times 11{,}4) = 7{,}17.$$

$$a_1 = 9{,}56 - (-0{,}97\times(3-3{,}2) + 0{,}48\times(9-11{,}4)) = -8{,}21,$$

$$a_2 = -6{,}44,\ a_3 = 0{,}63,\ a_4 = 4{,}93,\ a_5 = 9{,}08,$$

$c1$ = -0,98, c_2 = -0,40, c_3 = 1,38.

Сумма квадратов остатков равна RSS=0,6602. Оценка s^2:

s^2= 0,6602/(15-9) =0,11.

Для проверки значимости индивидуальных и временных эффектов

а) H_0: $a_1 = a_2 = ... = a_{N-1} = 0$, $\gamma_1 = \gamma_2 = ... = \gamma_{T-1} = 0$ используют F-критерий:

$$F(N+T-2, NT-N-T-K+1) = \frac{(RSS^r - RSS)/(N+T-2)}{RSS/(NT-N-T-K+1)},$$

где RSS^r – сумма квадратов остатков обычной регрессии $y_{it} = a + \beta' x_{it} + \varepsilon_{it}$.

Для тестирования на временные эффекты

б) H_0: $\gamma_1 = \gamma_2 = ... = \gamma_{T-1} = 0$ найдем

$$F(T-1, NT-N-T-K+1) = \frac{(RSS^r - RSS)/(T-1)}{RSS/(NT-N-T-K+1)},$$

где RSS^r – сумма квадратов остатков регрессии с фиксированными эффектами $y_{it} = a_i + \beta' x_{it} + \varepsilon_{it}$.

Для тестирования на индивидуальные эффекты

в) H_0: $a_1=a_2=...=a_{N-1}=0$ вычислим

$$F(N-1, NT-N-T-K+1) = \frac{(RSS^r - RSS)/(N-1)}{RSS/(NT-N-T-K+1)},$$

где RSS^r – сумма квадратов остатков регрессии с фиксированными временными эффектами $y_{it} = \gamma_t + \beta' x_{it} + \varepsilon_{it}$.

Для проверки гипотезы о значимости коэффициентов при факторах

г) H_0: $\beta_1=\beta_2=...=\beta_K=0$

$$F(K, NT-N-T-K+1) = \frac{(TSS - RSS)/K}{RSS/(NT-N-T-K+1)},$$

где $TSS = \sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \bar{y}_i - \bar{y}_t + \bar{\bar{y}})^2$.

Для данных примера 1 получим:

а) RSS=0,6602, RSS^r= 104,415, $F_{набл}$=157,15, $F_{крит}(6,6,\alpha=0,01)$ = 8,46. Таким образом, гипотеза о равенстве нулю одновременно индивидуальных и временных эффектов отвергается при уровне значимости α=0,01.

Покажем, как можно проверить значимость только фиксированных и только временных эффектов.

б) Для проверки значимости временных эффектов найдем суммы квадратов остатков двунаправленной модели и однонаправленной модели. Сумма квадратов остатков для двунаправленной модели равна RSS=0,6602. Сумма квадратов однонаправленной модели RSS^r= 0,6659 (см. раздел 2.2). По этим данным вычислим $F_{набл}$= 0,03, $F_{крит}(2,6,\alpha=0,01)$ = 10,92. Наблюдаемое значение не попадает в критическую область, следовательно, временные эффекты не значимы.

в) Для проверки значимости индивидуальных эффектов необходимо найти сумму квадратов остатков двунаправленной модели и модели с фиксированными временными эффектами $y_{it} = \gamma_t + \beta' x_{it} + \varepsilon_{it}$. Получим RSS=0,6602, RSS^r= 91,919, $F_{набл}$=207,3435, $F_{крит}(4,6,\alpha=0,01)$ = 9,15. Таким образом, индивидуальные эффекты значимо отличается от нуля с надежностью γ=1-α = 0,99.

г) Для проверки значимости коэффициентов регрессии при факторах найдем RSS=0,6602, TSS= 13,936, тогда $F_{набл}$=60,32, $F_{крит}(2,6,\alpha=0,01)$ = 10,92. Наблюдаемые значения попадают в критическую область. Таким образом, оценки коэффициентов регрессии при факторах значимо отличаются от нуля с надежностью γ=1-α = 0,99.

2.5. Незакрытые панели и модель с фиксированными эффектами

Проблема отсутствующих наблюдений возникает при анализе панельных данных достаточно часто. Часть респондентов по тем или иным причинам временно отсутствует или выбывает и их приходится заменять другими, фирмы сливаются, разоряются или поглощают друг друга и т.п. В этом случае панель называют незакрытой.

Расчетные формулы для модели с фиксированными эффектами следует модифицировать так, чтобы учесть разное число наблюдений. Пусть T_i – число наблюдений для i-го объекта. Тогда полный размер выборки равен $\sum_{i=1}^{N} T_i$. При исчислении групповых средних необходимо использовать соответствующие размеры групп T_i:

$$\bar{y}_i = \frac{1}{T_i}\sum_{t=1}^{T_i} y_{it}\,,\ \bar{x}_i = \frac{1}{T_i}\sum_{t=1}^{T_i} x_{it}\,.$$

Матрица перекрестных произведений $S_{xx}^{w} = \mathrm{X}'\mathbf{M}_d\mathrm{X}$ заменяется на сумму матриц сумм квадратов и перекрестных произведений:

$$S_{xx}^{w} = \sum_{i=1}^{N}(\sum_{t=1}^{T_i}(\mathrm{x}_{it} - \bar{\mathrm{x}}_i)(\mathrm{x}_{it} - \bar{\mathrm{x}}_i)').$$

Аналогично и для S_{xy}^{w}: $S_{xy}^{w} = \sum_{i=1}^{N}\sum_{t=1}^{T_i}(x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)$.

Оценки коэффициентов регрессии получаются из уравнения $\boldsymbol{b}^{w} = [S_{xx}^{w}]^{-1} S_{xy}^{w}$. *Within* - оценка по-прежнему совпадает с МНК – оценкой, вычисленной по отклонениям от средних:

$$(y_{it} - \bar{y}_i) = (x_{it} - \bar{x}_i)'\beta + u_{it}.$$

3. МОДЕЛЬ СО СЛУЧАЙНЫМИ ЭФФЕКТАМИ

Иногда есть основания предполагать, что индивидуальные эффекты не коррелированны с регрессорами. Например, если данные являются случайной выборкой из большой популяции. Тогда индивидуальные эффекты можно рассматривать как одну из составляющих ошибки.

Переформулируем модель, включающую K регрессоров, в виде:

$$y_{it} = \alpha + \boldsymbol{\beta}\,\mathbf{x}_{it} + u_i + \varepsilon_{it}. \qquad (18)$$

Компонента u_i является случайным отклонением (случайной ошибкой), соответствующей i-му объекту и постоянной во времени. Эта величина может соответствовать, например, суммарному влиянию факторов, специфических для отдельной фирмы, семьи, индивидуума и т.п. и не включенных в число регрессоров. Допустим, что

$$\begin{gathered}
\mathrm{E}[\varepsilon_{it}] = \mathrm{E}[u_i] = 0,\\
\mathrm{E}[\varepsilon_{it}^2] = \sigma_\varepsilon^2,\\
\mathrm{E}[u_i^2] = \sigma_u^2,\\
\mathrm{E}[\varepsilon_{it}\,u_j] = 0 \text{ для всех } i, t \text{ и } j, \qquad (19)\\
\mathrm{E}[\varepsilon_{it}\varepsilon_{js}] = 0 \text{ если } t \neq s \text{ или } i \neq j,\\
\mathrm{E}[u_i u_j] = 0 \text{ если } i \neq j.
\end{gathered}$$

Рассмотрим T наблюдений, соответствующих i-му объекту. Обозначим суммарную ошибку w_{it}

$$w_{it} = \varepsilon_{it} + u_i$$

и

$$\mathbf{w}_i = [w_{i1}, w_{t2}, \ldots, w_{iT}]'.$$

Тогда

$$\mathrm{E}[w_{it}^2] = \sigma_\varepsilon^2 + \sigma_u^2,$$

$$E[w_{it}\, w_{is}] = \sigma_u^2,\ t \neq s.$$

Ковариационная матрица наблюдений для i-го объекта $\Omega = E[\mathbf{w}_i\, \mathbf{w}'_i]$ равна:

$$\Omega = \begin{bmatrix} \sigma_\varepsilon^2+\sigma_u^2 & \sigma_u^2 & \sigma_u^2 & \Lambda & \sigma_u^2 \\ \sigma_u^2 & \sigma_\varepsilon^2+\sigma_u^2 & \sigma_u^2 & \Lambda & \sigma_u^2 \\ & & \mathrm{M} & & \\ \sigma_u^2 & \sigma_u^2 & \sigma_u^2 & & \sigma_\varepsilon^2+\sigma_u^2 \end{bmatrix} = \sigma_\varepsilon^2 \mathbf{I} + \sigma_u^2 \mathbf{i}\mathbf{i}' \qquad (20)$$

где $\mathbf{I}$ – единичная матрица, $\mathbf{i}$ – единичный вектор размерности $T \times 1$.

Так как ошибки для объектов i и j независимы, ковариационная матрица всех NT наблюдений будет иметь вид:

$$V = \begin{bmatrix} \Omega & 0 & 0 & \Lambda & 0 \\ 0 & \Omega & 0 & \Lambda & 0 \\ & & & \mathrm{M} & 0 \\ 0 & 0 & 0 & \Lambda & \Omega \end{bmatrix}. \qquad (21)$$

Таким образом, модель случайных эффектов соответствует модели линейной регрессии при гетероскедастичности ошибок.

В этом случае можно применять как МНК – оценки, так и оценки модели с фиксированными эффектами или *between*-оценки. Они будут несмещенными и состоятельными, но неэффективными.

Эффективными, как известно, являются оценки обобщенного метода наименьших квадратов (*GLS* - оценки). Для уравнения регрессии $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ *GLS* - оценки рассчитываются по формуле:

$$\hat{\beta}_{GLS} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y},$$

$$Var(\hat{\beta}_{GLS} \mid X) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}.$$

GLS-оценки можно также найти из преобразованного уравнения:

$$\mathbf{V}^{-1/2}\mathbf{Y} = \mathbf{V}^{-1/2}\mathbf{X}\beta + \nu.$$

3.1. Обобщенный метод наименьших квадратов

Для использования обобщенного метода наименьших квадратов (GLS), необходимо найти матрицу $\mathbf{V}^{-1/2} = \mathbf{I} \otimes \Omega^{-1/2}$, где $\otimes$ - символ произведения Кронекера. Следовательно, нам необходимо найти матрицу $\Omega^{-1/2}$. Матрица Ω имеет достаточно простую структуру. Если допустить, что дисперсии σ_ε^2 и σ_u^2 известны, можно выписать явное выражение для $\mathbf{\Omega}^{-1/2}$:

$$\Omega^{-1/2} = \mathrm{I} - \frac{\theta}{T}\mathbf{i}\mathbf{i}',$$

где θ - параметр, зависящий от σ_ε^2 и σ_u^2:

$$\theta = 1 - \frac{\sigma_\varepsilon}{\sqrt{T\sigma_u^2 + \sigma_\varepsilon^2}}.$$

Умножение y_i на $\mathbf{\Omega}^{-1/2}$ означает следующее преобразование:

$$y_i^* = \Omega^{-1/2} y_i = \begin{bmatrix} y_{i1} - \theta\bar{y}_i \\ y_{i2} - \theta\bar{y}_i \\ \mathrm{M} \\ y_{iT} - \theta\bar{y}_i \end{bmatrix}, \qquad (22)$$

и аналогично для $\mathbf{x}_i$:

$$x_i^* = \Omega^{-1/2} x_i = \begin{bmatrix} x_{i1} - \theta\bar{x}_i \\ x_{i2} - \theta\bar{x}_i \\ \mathrm{M} \\ x_{iT} - \theta\bar{x}_i \end{bmatrix}.$$

Таким образом, оценки обобщенного метода наименьших квадратов можно найти, рассчитав коэффициенты регрессии преобразованных переменных y_{it} на аналогичным способом преобразованные $\mathbf{x}_{it}$. Заметим, что эта процедура имеет определенное сходство с преобразованиями в

модели с фиксированными эффектами, когда полагают $\theta = 1$. (Условие $\theta = 1$ можно интерпретировать как то, что дисперсия σ_ϵ равна нулю, следовательно, ошибки совпадают с u_i. В этом случае модели с фиксированными и случайными эффектами совпадают).

Можно показать, что оценки обобщенного метода наименьших квадратов, подобно МНК оценкам, могут быть вычислены как матричные взвешенные межгрупповых и внутригрупповых:

$$\hat{\boldsymbol{\beta}} = \hat{\mathbf{F}}^{\mathbf{w}}\mathbf{b}^{\mathbf{w}} + (\mathbf{I} - \hat{\mathbf{F}}^{\mathbf{w}})\mathbf{b}^{\mathbf{b}}, \qquad (23)$$

где

$$\hat{F}^{w} = \left[S_{xx}^{w} + \lambda S_{xx}^{b}\right]^{-1} \cdot S_{xx}^{w},$$

$$\lambda = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + T\sigma_u^2} = (1-\theta)^2$$

Формула (23) позволяет проанализировать, к чему сводятся *GLS*-оценки в зависимости от параметра λ.

Если $\lambda = 1$, то *GLS*-оценки совпадают с оценками простой регрессии. Из формулы (23) видно, что это возможно, когда σ_u^2 равно нулю, то есть индивидуальных эффектов вовсе не существует. Но, если $\lambda \neq 1$, то МНК – оценки становятся неэффективными. По сравнению с обобщенным, обычный метод наименьших квадратов придает слишком большой вес вариации между объектами. Он объясняет ее целиком изменениями независимых переменных, вместо того, чтобы допустить, что некоторые колебания значений признака объясняется случайной ошибкой u_i.

Если $\lambda = 0$ получим *within* - оценки (фиксированных эффектов). Если число периодов наблюдения T конечно то, чтобы параметр λ равнялся нулю, необходимо чтобы дисперсия σ_ε^2 была равна нулю, то есть все различия между объектами объяснялись случайными величинами u_i, которые постоянны во времени. Другой возможный случай, когда число

периодов наблюдения стремится к бесконечности $T \to \infty$, то $\lambda \to 0$ и оценка модели со случайными эффектами стремится к оценке фиксированных эффектов.

Пример 1 (продолжение).

Для иллюстрации допустим, что $\sigma^2_u = 10$, $\sigma^2_\varepsilon = 1$. Тогда

$$\theta = 1 - \frac{\sigma_\epsilon}{\sqrt{T\sigma_u^2 + \sigma_\epsilon^2}} = 1 - \frac{1}{\sqrt{3\times 10 + 1}} = 1 - 0{,}1796 = 0{,}8204,$$

$$\lambda = (1-\theta)^2 = 0{,}0323,$$

$$\hat{\mathbf{F}}^{\mathbf{w}} = \left[\begin{bmatrix} 24 & 21 \\ 21 & 50 \end{bmatrix} + 0{,}0323 \begin{bmatrix} 68{,}4 & -7{,}2 \\ -7{,}2 & 135{,}6 \end{bmatrix}\right]^{-1} \begin{bmatrix} 24 & 21 \\ 21 & 50 \end{bmatrix} = \begin{bmatrix} 0{,}8744 & 0{,}1041 \\ 0{,}0522 & 0{,}8798 \end{bmatrix}.$$

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 0{,}8744 & 0{,}1041 \\ 0{,}0522 & 0{,}8798 \end{bmatrix} \begin{bmatrix} -0{,}97 \\ 0{,}49 \end{bmatrix} + \begin{bmatrix} 0{,}1256 & -0{,}1041 \\ -0{,}0522 & 0{,}1202 \end{bmatrix} \begin{bmatrix} -0{,}41 \\ 1{,}62 \end{bmatrix} = \begin{bmatrix} -1{,}0185 \\ 0{,}5969 \end{bmatrix}.$$

Подобные расчеты легко провести для разных значений параметра θ. На рис.6 показано изменение значений коэффициентов, если параметр меняется от 0 до 1.

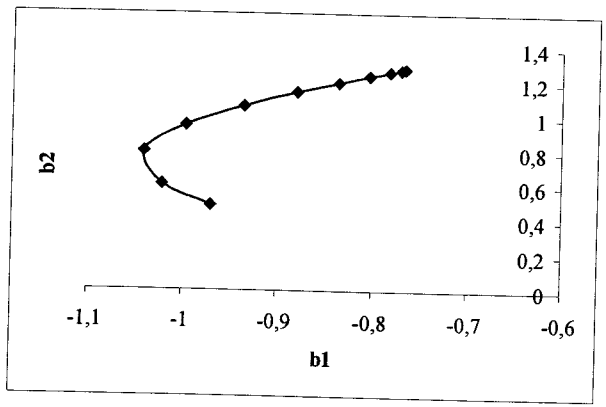

Рис.6. Изменение коэффициентов b_1 и b_2 в зависимости от параметра θ

К сожалению дисперсии индивидуальных эффектов и случайных членов почти никогда заранее неизвестны. Соответственно невозможно вычислить точные значения параметров θ или λ и матрицы $\Omega^{-1/2}$.

3.2. Доступный метод наименьших квадратов для модели со случайными эффектами

Так как дисперсии компонент ошибок неизвестны, необходимо попытаться оценить их по выборке, а затем воспользоваться доступным обобщенным методом наименьших квадратов (FGLS).

Существует несколько эвристических подход к оценке неизвестных компонент дисперсии $\hat{\sigma}_u^2$ и $\hat{\sigma}_\varepsilon^2$.

Рассмотрим один из них. Запишем исходное уравнение

$$y_{it} = \alpha + \beta' \mathbf{x}_{it} + u_i + \varepsilon_{it}$$

и уравнение для групповых средних

$$\bar{y}_i = \alpha + \beta' \bar{\mathbf{x}}_i + u_i + \bar{\varepsilon}_i. \qquad (24)$$

Вычитание из каждого уравнения среднего по группе позволяет избавиться от неоднородности:

$$y_{it} - \bar{y}_i = \beta' (\mathbf{x}_{it} - \bar{\mathbf{x}}_i) + (\varepsilon_{it} - \bar{\varepsilon}_i). \qquad (25)$$

Это модель *within* – регрессии, и она не содержит значений индивидуальных эффектов u_i. Следовательно, оценку $\hat{\sigma}_\varepsilon^2$ можно найти на основании остатков модели с фиксированными эффектами:

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - a_i - \mathbf{x}'_{it}\mathbf{b})^2}{NT - N - K}$$

Рассмотрим далее средние ошибки для каждой группы:

$$\varepsilon_{**_i} = \bar{y}_i - a - \beta' \bar{\mathbf{x}}_i = \bar{\varepsilon}_i - u_i, \quad i = 1,2,...N. \qquad (26)$$

Средние ошибки для объектов взаимно независимы, их дисперсия равна:

$$D(\varepsilon_{**_i}) = \sigma_{**}^2 = \frac{\sigma_\varepsilon^2}{T} + \sigma_u^2. \qquad (27)$$

Несмещенную оценку для величины $\sigma_\varepsilon^2/T + \sigma_u^2$ можно найти из *between* – регрессии:

$$\hat{\sigma}_{**}^2 = \frac{e'_{**}e_{**}}{n-k}. \qquad (28)$$

где e_{**} - остатки *between* – регрессии.

Это приводит к следующей оценке дисперсии

$$\hat{\sigma}_u^2 = \hat{\sigma}_{**}^2 - \frac{\hat{\sigma}_\varepsilon^2}{T}. \qquad (29)$$

Оценка дисперсии (29) является несмещенной, но на практике может быть отрицательной, что необходимо учитывать.

Пример 1 (продолжение).

Сумма квадратов остатков для модели с фиксированными эффектами равна 0,665982. Следовательно, оценка остаточной дисперсии равна:

$$\hat{\sigma}_\varepsilon^2 = \frac{0{,}665982}{8} = 0{,}083248.$$

Групповые средние приведены в таблице.

$\bar{x}1$	$\bar{x}2$	$\bar{y}$	Прогноз	Остатки
3	9	2,7	5,74	-3,04
5	7	3,5	1,66	1,84
0	12	13	11,86	1,14
2	14	14,4	14,28	0,12
6	15	14,2	14,25	-0,05

Сумма квадратов остатков равна 13,94316. Следовательно,

$$\frac{\hat{\sigma}_\varepsilon^2}{T} + \hat{\sigma}_u^2 = \frac{13{,}94316}{2} = 6{,}971577,$$

$$\hat{\sigma}_u^2 = 6{,}971577 - \frac{0{,}083248}{3} = 6{,}943826$$

Найдем оценку параметра θ:

$$\hat{\theta} = 1 - \frac{\sqrt{0{,}083248}}{\sqrt{0.083248 + 3 \times 6{,}943826}} = 0{,}93961,$$

$$\lambda = (1-\theta)^2 = 0{,}00398.$$

$$\hat{\mathbf{F}}^{\mathbf{w}} = \begin{bmatrix} 24 & 21 \\ 21 & 50 \end{bmatrix} + 0{,}00398 \begin{bmatrix} 68{,}4 & -7{,}2 \\ -7{,}2 & 135{,}6 \end{bmatrix}^{-1} \begin{bmatrix} 24 & 21 \\ 21 & 50 \end{bmatrix} = \begin{bmatrix} 0{,}98175 & 0{,}01623 \\ 0{,}00814 & 0{,}98259 \end{bmatrix}$$

$$\mathbf{b} = \begin{bmatrix} 0{,}98175 & 0{,}01623 \\ 0{,}00814 & 0{,}98259 \end{bmatrix} \begin{bmatrix} -0{,}97 \\ 0{,}49 \end{bmatrix} + \begin{bmatrix} 0{,}01825 & -0{,}01623 \\ -0{,}00814 & 0{,}01741 \end{bmatrix} \begin{bmatrix} -0{,}41 \\ 1{,}62 \end{bmatrix} = \begin{bmatrix} -0{,}97814 \\ 0{,}50460 \end{bmatrix}$$

Используем доступный метод наименьших квадратов для вычисления вектора **b.** Таблица данных после *GLS* – преобразования выглядит следующим образом.

i	t	$x1_{it} - \theta\,\bar{x}1_i$	$x2_{it} - \theta\,\bar{x}2_i$	$y_{it} - \theta\,\bar{y}_i$
1	1	0.1893	1.5678	0.7703
1	2	1.1893	1.5678	-0.6297
1	3	-0.8107	-1.4322	0.3703
2	1	0.3155	0.4416	0.0208
2	2	-0.6845	-0.5584	1.0208
2	3	1.3155	1.4416	-0.3792
3	1	0.0000	-0.2429	0.7202
3	2	0.0000	0.7571	0.6202
3	3	0.0000	1.7571	1.1202
4	1	-0.8738	-1.1167	0.8085
4	2	2.1262	-0.1167	-1.4915
4	3	-0.8738	3.8833	3.4085
5	1	-1.6215	-2.0536	1.0959
5	2	-0.6215	-0.0536	1.4959
5	3	3.3785	4.9464	0.0959

Результаты регрессии приведены в таблице.

	Коэффициенты регрессии	Стандартные ошибки коэффициентов	z-статистики	Границы 95 % доверительных интервалов	
a	6.9376	1.5956	4.35	3.810282	10.06497
b_1	-0.9781	0.0916	-10.67	-1.15775	-0.79853
b_2	0.5046	0.0635	7.95	0.380125	0.629071

3.3. Проверка значимости случайных эффектов

Бреуш и Паган (Breusch, Pagan) предложили метод множителей Лагранжа для проверки гипотезы о значимости случайных эффектов, основанный на остатках простой МНК - регрессии. Для проверки гипотезы

$$H_0 : \sigma_u^2 = 0,$$

$$H_1 : \sigma_u^2 \neq 0,$$

используется тестовая статистика:

$$\mathrm{LM} = \frac{NT}{2(T-1)} \left[\frac{\sum_{i=1}^{N} \left[\sum_{t=1}^{T} e_{it} \right]^2}{\sum_{i=1}^{N} \sum_{t=1}^{T} e_{it}^2} - 1 \right]^2$$

$$= \frac{NT}{2(T-1)} \left[\frac{\sum_{i=1}^{N} (T\bar{e}_i)^2}{\sum_{i=1}^{n} \sum_{t=1}^{T} e_{it}^2} - 1 \right]^2, \qquad (30)$$

где e_{it} – остатки в стандартной регрессионной модели.

При нулевой гипотезе LM подчиняется закону распределения Хи-квадрат с одной степенью свободы. Выражение (30) можно более компактно записать в матричной записи. Пусть **D** - матрица фиктивных

переменных, определенная в (2) и пусть **e** - вектор остатков, полученный методом наименьших квадратов. Тогда

$$\mathrm{LM} = \frac{nT}{2(T-1)}\left[\frac{\mathbf{e}'\mathbf{D}\mathbf{D}'\mathbf{e}}{\mathbf{e}'\mathbf{e}} - 1\right]^2. \qquad (31)$$

Пример 1 (продолжение).

В таблице приведены значения остатков для уравнения регрессии

$$\hat{y}_{it} = -2{,}61 - 0{,}77x1 + 1{,}28\,x2$$

Рассчитаем сумму квадратов остатков и сумму квадратов их средних по группам:

i	t	e_{it}	e^2_{it}	$T\bar{e}_i$	$(T\bar{e}_i)^2$
1	1	-4,61711	21,31772		
1	2	-5,25105	27,57349		
1	3	-1,93439	3,741861	-11,8025	139,3001
2	1	0,763807	0,583401		
2	2	2,280671	5,20146		
2	3	-0,15306	0,023427	2,89142	8,360312
3	1	1,401763	1,96494		
3	2	0,018834	0,000355		
3	3	-0,7641	0,583842	0,656501	0,430994
4	1	2,284899	5,220763		
4	2	1,000165	1,00033		
4	3	-1,52975	2,340129	1,755316	3,081134
5	1	4,683094	21,93137		
5	2	3,283301	10,78007		
5	3	-1,46709	2,152338	6,49931	42,24103
Суммы квадратов			104,4155		193,4136

Найдем значение LM:

$$LM = \frac{5\times 3}{2\times(3-1)}\left[\frac{104{,}41}{193{,}41} - 1\right]^2 = 2{,}724.$$

На 5-ти процентном уровне значимости критическое значение хи-квадрат с одной степенью свободы составляет 3,842. Таким образом,

наблюдаемое значение не попадает в критическую область. Следовательно, нет оснований предпочесть модель со случайными эффектами простой регрессии.

3.4. Тест Хаусмана для сравнения моделей с фиксированными и случайными эффектами

Мы подробно рассмотрели модели с фиксированными и случайными эффектами. Возникает вопрос: какую из них следует использовать?

Если анализируемая совокупность охватывает страны, регионы, крупные предприятия или отрасли промышленности, то интересно получить прогноз для конкретной страны, предприятия или региона. Каждый такой объект уникален в своем роде, имеет свои собственные особенности, влияние которых учитывается с помощью параметров a_i. Подход фиксированных эффектов позволяет построить оценки при условии этих параметров. Но, если мы захотим распространить модель для объектов, не вошедших в выборку, то значения индивидуальных эффектов для них неизвестны и модель не применима.

Если же анализируется случайная выборка из большой популяции, то величины индивидуальных эффектов не представляют интереса. Важно описать поведение совокупности в целом и построить прогнозы для типичных представителей совокупности. Поэтому, если имеющиеся данные извлечены из большой популяции, то возможно лучший вариант – случайные эффекты.

Если выборка исчерпывает совокупность, как в случае регионов, отраслей промышленности, то естественным кандидатом является модель с фиксированными эффектами.

С другой стороны, необходимо учитывать свойства оценок. Модель фиксированных эффектов гарантировано позволяет при сделанных предположениях получить несмещенные и состоятельные оценки. Но, если число периодов наблюдения невелико, очень важно получить эффективные оценки. Пренебрегать такой возможностью никогда не следует. Оценки модели со случайным эффектами являются оптимальными, когда индивидуальные эффекты не коррелированны с регрессорами, но в противном случае они не будут состоятельными.

Если объем выборки велик, интуитивно более привлекательна модель со случайными эффектами. Модель с фиксированными эффектами просто вычисляет индивидуальные эффекты, реализованные в данной выборке, которые сами по себе не представляют интереса. Но метод фиксированных эффектов не требует некоррелированности индивидуальных эффектов и остальных регрессоров, что по умолчанию принимается в модели со случайными эффектами. С другой стороны, если корреляция отсутствует, использование оценок фиксированных эффектов приведет к потере степеней свободы.

Можно приводить различные аргументы, что особенности сбора, выборочные характеристики данных, или иные известные факторы говорят в пользу модели с фиксированными или случайными эффектами. Тем не менее априорно доказать возможность применения модели со случайными эффектами удается редко.

Существует способ статистической проверки гипотезы, ортогональны ли случайные эффекты и регрессоры, предложенный Хаусманном (***Hausman Test***). Его подход основан на том, что если гипотеза об отсутствии корреляции верна, то оценки моделей с фиксированными и случайными эффектами являются состоятельными, но оценки случайных эффективны. Поэтому при выполнении нулевой гипотезы между оценками нет систематического смещения. При

альтернативной гипотезе состоятельны лишь оценки модели с фиксированными эффектами.

При выполнении нулевой гипотезы статистика

$$\mathbf{W} = [b_{FE} - b_{RE}]'[Cov(b_{FE}) - Cov(b_{RE})]^{-1}[b_{FE} - b_{RE}], \qquad (32)$$

асимптотически подчиняется закону распределения Хи-квадрат с K степенями свободы, где $Cov(b_{FE})$ и $Cov(b_{FE})$ - оценки ковариационных матриц для параметров моделей с фиксированными и случайными эффектами.

Если наблюдаемое значение статистки **W** не попадает в критическую область $W_{набл} < \chi^2_{крит}$, то различия между оценками не являются систематическими. Это означает, что можно выбрать модель со случайными эффектами. В противном случае, когда $W_{набл} > \chi^2_{крит}$, следует выбрать модель с фиксированными эффектами.

Для данных примера 1 оценки случайных и фиксированных эффектов достаточно близки.

Параметры	Оценки модели фиксированных эффектов	Оценки модели случайных эффектов
β_1	-0.96983	-0.97814
β_2	0.489328	0.504598

Тест Хаусмана ($W_{набл.}$=0,001) позволяет сделать вывод в пользу модели со случайными эффектами. Но, следует иметь в виду, что он является асимптотическим, то есть требует достаточного количества объектов.

ЗАКЛЮЧЕНИЕ

Изложенные в настоящем учебном пособии модели и методы, разумеется, не исчерпывают всего спектра вопросов, с которыми сталкивается исследователь при анализе панельных данных.

Методология анализа панельных данных интенсивно развивается. Исследованы свойства более общих спецификаций, предложены методы оценивания для моделей с дискретными зависимыми переменными, динамических моделей и т.д. Накоплен значительный опыт использования разработанных методов для широкого спектра экономических задач. Обзор современных достижений содержится в [4,6].

Авторы надеются, что данное пособие может послужить полезным введением в данную проблематику для студентов и аспирантов при изучении курсов многомерного статистического анализа и эконометрики.

Приложение 1. Краткий обзор команд обработки панельных данных пакета Stata

Пакет Stata является мощным пакетом статистического анализа. Он позволяет осуществлять разнообразные манипуляции с данными и включает множество способов их обработки. В том числе пакет включает основные средства анализа панельных данных. Мы опишем лишь команды, позволяющие оценивать модели с фиксированными и случайными эффектами.

При анализе панельных данных предполагается, что один из признаков содержит номера или коды-идентификаторы объектов. Например, набор данных содержит 5 переменных. Переменная **id** – содержит номера объектов, **time** – номер периода (год), **x1, x2, y** – анализируемые данные.

id	time	x1	x2	y
10	1991	3	10	3.3
10	1992	4	10	1.9
10	1993	2	7	2.9
11	1991	5	7	3.3
11	1992	4	6	4.3
11	1993	6	8	2.9
24	1991	0	11	12.9
24	1992	0	12	12.8
24	1993	0	13	13.3
47	1991	1	12	14.3
47	1992	4	13	12
47	1993	1	17	16.9
56	1991	4	12	14.4
56	1992	5	14	14.8
56	1993	9	19	13.4

Данные могут быть введены вручную или скопированы в окно редактора из других приложений с помощью стандартных возможностей Windows - команд Copy / Paste.

Обработка данных осуществляется с помощью команд, вводимых с клавиатуры. Для расчета параметров моделей с фиксированными и случайными эффектами служат команды: **iis, xtreg , xttest0, xthausman.** Синтаксис их приведен ниже.

iis *varname* – указывает переменную, содержащую номера объектов;

xtreg *depvar varlist*, **fe** – расчет модели с фиксированными эффектами;

xtreg *depvar varlist*, **re** – расчет модели со случайными эффектами;

xttest0 - расчет LM–теста для модели со случайными эффектами;

xthausman - расчет теста Хаусмана;

где *depvar* - имя зависимой переменной; *varlist* – список независимых переменных.

Например,

iis id
xtreg y x1 x2, fe
xtreg y x1 x2, re
xttest0
xthauman

Результат выполнения команд:

iis id
xtreg y x1 x2, fe

```
Fixed-effects (within) regression               Number of obs      =        15
Group variable (i) : id                         Number of groups   =         5

R-sq:  within  = 0.9564                         Obs per group: min =         3
       between = 0.5026                                        avg =       3.0
       overall = 0.5130                                        max =         3

                                                F(2,8)             =     87.77
corr(u_i, Xb)  = 0.2709                         Prob > F           =    0.0000

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |  -.9698287   .0740543   -13.10   0.000    -1.140598   -.7990592
          x2 |    .489328   .0513063     9.54   0.000     .3710155    .6076406
       _cons |   7.085112   .4855476    14.59   0.000     5.965438    8.204787
-------------+----------------------------------------------------------------
     sigma_u |  4.3678799
     sigma_e |  .28852672
         rho |   .9956555   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(4, 8) =   311.57                 Prob > F = 0.0000
```


xtreg y x1 x2, re

```
Random-effects GLS regression                   Number of obs      =        15
Group variable (i) : id                         Number of groups   =         5

R-sq:  within  = 0.9560                         Obs per group: min =         3
       between = 0.5113                                        avg =       3.0
       overall = 0.5213                                        max =         3

Random effects u_i ~ Gaussian                   Wald chi2(2)       =    117.69
corr(u_i, X)       = 0 (assumed)                Prob > chi2        =    0.0000

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |  -.9781385   .0916405   -10.67   0.000    -1.157751   -.7985264
          x2 |   .5045978   .0635076     7.95   0.000     .3801251    .6290705
       _cons |   6.937628   1.595614     4.35   0.000     3.810282    10.06497
-------------+----------------------------------------------------------------
     sigma_u |  2.6351144
     sigma_e |  .28852672
         rho |   .9881533   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

xttest0

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

        y[id,t] = Xb + u[id] + e[id,t]

        Estimated results:
                         |       Var     sd = sqrt(Var)
                 --------+------------------------------
                       y |      31.214        5.586949
                       e |    .0832477        .2885267
                       u |    6.943828        2.635114

        Test:   Var(u) = 0
                                  chi2(1) =     2.72
                               Prob > chi2 =     0.0988
```

xthaus

```
Hausman specification test

                 ---- Coefficients ----
              |      Fixed       Random
            y |    Effects      Effects        Difference
--------------+----------------------------------------------
           x1 |  -.9698287    -.9781385         .0083097
           x2 |    .489328     .5045978        -.0152698

    Test:  Ho:  difference in coefficients not systematic

                chi2(  2) = (b-B)'[S^(-1)](b-B), S = (S_fe - S_re)
                          =      0.00
               Prob>chi2 =      1.0000
```

Для самостоятельной работы с пакетом понадобится изучить и другие команды. Описание всех возможностей пакета далеко выходит за рамки данного пособия.

Приведем лишь пример команд для расчета *within* и *between* – преобразований и расчета регрессий для данных примера.

```
/* within */
sort id
by id: egen x1m=mean(x1)
by id: egen x2m=mean(x2)
by id: egen ym=mean(y)
gen yw=y-ym
gen x1w=x1-x1m
gen x2w=x2-x2m
reg  yw x1w x2w

/* between */
sort time
by time: egen x1mt=mean(x1)
by time: egen x2mt=mean(x2)
by time: egen ymt=mean(y)
gen yb=y-ymt
gen x1b=x1-x1mt
gen x2b=x2-x2mt
reg  yb x1b x2b
```


Приложение 2. Варианты заданий и исходные данные для самостоятельной работы на ЭВМ

В таблице 3 приведены значения некоторых показателей социально-экономического развития следующих европейских стран: 1 – Австрия; 2 – Бельгия; 3 – Дания; 4 - Финляндия; 5 – Франция; 6 – Греция; 7 – Исландия; 8 - Ирландия; 9 – Италия; 10 – Голландия; 11 – Норвегия; 12 – Португалия; 13 – Испания; 14 – Швеция; 15 - Швейцария; 16 – Великобритания.

Рассматриваются следующие показатели:

x_1 - численность населения, млн. чел.;

x_2 - численность занятых в экономике, млн. чел.;

x_3 - инвестиции, млрд. долларов США;

x_4 - физический капитал, рассчитанный как накопленные инвестиции за 1960-1995 годы при норме амортизации 5% годовых;

x_5 - кредиты, предоставленные банками, % ВВП;

x_6 - расходы на 1 студента высшего учебного заведения, % ВВП на душу населения;

x_7 - численность студентов высших и средних специальных учебных заведений, млн. чел.;

x_8 - ВВП, млрд. долларов США;

x_9 - индекс потребительских цен, %;

x_{10} - потребление электроэнергии, квт-час. на душу населения;

x_{11} - частное потребление, млрд. долларов США;

x_{12} - коэффициент демографической нагрузки (соотношение численности населения в нетрудоспособном и трудоспособном возрасте);

x_{13} - прирост национальных сбережений, млрд. долларов США.

Таблица 3

Таблица исходных данных

Year	*id*	*x1*	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*	*x8*	*x9*	*x10*	*x11*	*x12*	*x13*
1990	1	7.73	3.55	38.61	236	123.0	35.8	0.206	159.50	3.26	5587	89.1	0.48	40.0
1990	2	9.97	3.99	39.65	239	70.9	29.3	0.276	196.13	3.45	5817	124.7	0.50	40.7
1990	3	5.14	2.93	27.08	193	63.0	.	0.143	133.36	2.64	5650	65.4	0.48	29.4
1990	4	4.99	2.59	37.16	188	84.3	42.0	0.166	134.81	6.10	11822	70.6	0.48	30.5
1990	5	56.74	24.96	268.55	1672	106.1	23.4	1.699	1195.44	3.38	5321	712.3	0.52	254.1
1990	6	10.16	4.17	18.85	145	73.3	16.1	0.283	82.91	20.40	2802	60.7	0.49	14.9
1990	7	0.25	0.14	1.13	9	50.3	30.1	0.005	6.25	15.51	15345	3.8	0.55	1.0
1990	8	3.51	1.30	9.67	55	57.3	42.2	0.090	45.53	3.27	3385	26.5	0.63	9.6
1990	9	56.72	24.39	230.70	1354	90.1	.	1.452	1093.95	6.50	3784	670.7	0.45	213.0
1990	10	14.95	6.88	62.86	407	107.4	56.5	0.479	283.67	2.45	4917	166.5	0.45	72.5
1990	11	4.24	2.12	26.86	227	67.4	28.5	0.143	115.45	4.11	22824	57.0	0.54	29.5
1990	12	9.90	4.85	19.78	104	71.8	33.7	0.186	69.13	13.37	2379	43.6	0.51	19.6
1990	13	38.84	15.92	124.81	574	110.8	19.0	1.222	491.94	6.72	3239	307.1	0.50	106.5
1990	14	8.56	4.62	49.00	295	145.5	41.3	0.193	229.76	10.47	14061	117.0	0.55	41.1
1990	15	6.71	3.56	64.53	362	179.0	41.9	0.137	228.41	5.41	6997	130.9	0.45	73.2
1990	16	57.56	28.78	187.82	1155	123.0	41.9	1.258	975.51	9.48	4768	613.7	0.54	140.5
1991	1	7.83	3.60	41.82	266	122.5	36.2	0.217	166.65	3.33	5727	91.9	0.48	42.0
1991	2	10.00	4.10	38.02	265	68.3	.	.	201.15	3.21	6049	129.5	0.50	40.4
1991	3	5.15	2.94	25.59	209	66.0	41.7	0.150	134.08	2.40	5688	66.1	0.48	28.5
1991	4	5.01	2.56	24.87	204	94.6	53.1	0.174	121.38	4.12	11784	67.9	0.49	19.0
1991	5	57.06	25.10	258.41	1847	106.1	22.7	1.840	1201.01	3.22	5627	718.8	0.52	248.3
1991	6	10.25	4.20	20.87	158	66.2	16.2	0.272	89.05	19.47	2862	65.0	0.49	18.5
1991	7	0.26	0.14	1.28	10	51.4	.	0.006	6.73	6.81	14996	4.2	0.55	1.0
1991	8	3.53	1.34	8.91	61	52.4	39.3	0.101	46.19	3.19	3535	27.1	0.61	10.0
1991	9	56.75	24.40	236.73	1523	95.9	.	1.533	1150.70	6.30	3867	711.0	0.45	212.4
1991	10	15.07	6.93	61.94	448	107.5	54.3	0.494	290.20	3.13	5018	172.5	0.45	71.1
1991	11	4.26	2.13	25.22	241	62.0	29.5	0.154	117.76	3.42	23231	58.0	0.54	29.3
1991	12	9.87	4.84	20.93	120	77.0	35.1	0.191	78.32	11.35	2520	50.1	0.51	20.0
1991	13	38.92	15.96	129.96	675	108.8	19.2	1.302	528.60	5.93	3306	329.8	0.49	110.9
1991	14	8.62	4.65	42.86	323	143.1	42.8	0.207	239.33	9.34	14159	127.5	0.56	37.7
1991	15	6.80	3.60	59.12	403	176.9	42.5	0.143	232.68	5.84	7060	136.7	0.46	69.7
1991	16	57.81	28.90	163.70	1261	120.5	42.6	1.385	1012.16	5.85	4862	641.2	0.54	137.2
1992	1	7.91	3.64	44.74	298	122.7	33.7	0.221	187.21	4.03	5611	104.4	0.48	44.6
1992	2	10.05	4.12	42.16	294	149.4	.	0.307	224.84	2.43	6230	144.0	0.50	45.7
1992	3	5.17	2.90	26.64	226	58.9	.	0.164	147.09	2.10	5769	72.8	0.48	31.8
1992	4	5.04	2.57	18.33	212	92.3	55.3	0.188	106.44	2.60	11852	60.7	0.49	14.5
1992	5	57.37	25.24	260.01	2015	106.9	21.7	1.952	1322.21	2.37	5752	794.9	0.52	258.0
1992	6	10.32	4.34	20.58	171	68.3	.	0.299	98.45	15.87	2974	73.6	0.49	18.9
1992	7	0.26	0.15	1.20	11	53.7	.	0.006	6.91	3.96	14822	4.3	0.55	1.0
1992	8	3.55	1.35	8.42	66	55.8	39.2	0.108	52.37	3.12	3721	30.7	0.60	10.0
1992	9	56.86	24.45	237.84	1685	103.5	.	1.615	1219.15	5.08	3931	766.1	0.45	208.9
1992	10	15.18	6.98	66.23	492	107.4	.	0.507	321.93	3.18	5130	193.9	0.46	75.2
1992	11	4.29	2.19	26.09	255	63.6	28.9	0.166	126.31	2.34	23186	63.6	0.54	30.6
1992	12	9.87	4.83	24.47	138	81.0	29.7	0.248	94.51	8.94	2600	61.3	0.51	24.3
1992	13	39.01	16.38	130.67	772	105.9	18.8	1.371	577.31	5.92	3353	364.1	0.48	109.8
1992	14	8.67	4.68	40.86	348	139.2	46.8	0.223	247.56	2.29	13849	133.5	0.56	33.2

Year	*id*	*x1*	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*	*x8*	*x9*	*x10*	*x11*	*x12*	*x13*
1992	15	6.88	3.64	53.46	437	176.6	.	0.146	243.46	4.05	7028	145.5	0.46	68.6
1992	16	58.01	29.00	160.95	1359	118.2	45.9	1.528	1047.80	3.73	4852	669.5	0.54	133.5
1993	1	7.99	3.67	41.93	325	124.6	31.6	0.227	182.71	3.63	5600	102.7	0.48	41.3
1993	2	10.08	4.13	38.23	317	150.5	29.4	0.322	214.05	2.75	6277	136.3	0.50	45.2
1993	3	5.19	2.91	22.80	237	53.6	58.2	0.170	138.83	1.26	5838	69.4	0.48	30.3
1993	4	5.07	2.58	11.78	213	84.0	57.6	0.197	84.45	2.10	12291	48.2	0.49	11.8
1993	5	57.67	25.37	214.27	2128	103.7	24.4	2.083	1249.66	2.11	5762	760.9	0.53	224.9
1993	6	10.38	4.36	18.28	181	72.4	23.7	0.314	92.20	14.41	3004	69.4	0.49	16.6
1993	7	0.26	0.15	0.96	11	56.2	43.9	0.007	6.09	4.08	15474	3.7	0.55	1.0
1993	8	3.56	1.39	7.31	70	50.4	38.7	0.118	49.23	1.41	3802	27.8	0.58	9.7
1993	9	57.05	25.10	166.67	1767	104.3	22.7	1.770	985.15	4.48	3933	611.0	0.45	176.4
1993	10	15.28	7.03	58.07	526	109.9	46.1	0.532	313.07	2.58	5152	189.4	0.46	72.5
1993	11	4.31	2.20	25.07	267	59.1	48.5	0.177	116.11	2.27	23380	58.0	0.55	28.6
1993	12	9.88	4.84	19.42	151	84.4	26.1	0.277	83.73	6.80	2629	55.7	0.51	19.7
1993	13	39.08	16.41	95.16	829	102.5	17.5	1.469	478.96	4.57	3344	302.4	0.48	89.7
1993	14	8.72	4.71	24.66	355	139.3	.	0.234	185.81	4.64	13893	102.3	0.56	22.0
1993	15	6.94	3.68	48.93	464	179.0	44.9	0.149	236.73	3.32	6878	141.8	0.46	68.3
1993	16	58.19	29.10	141.91	1433	116.2	44.1	1.664	942.89	1.56	4917	606.3	0.54	118.8
1994	1	8.03	3.77	46.58	355	126.3	.	0.234	195.94	2.96	5659	109.9	0.48	44.6
1994	2	10.12	4.15	41.04	342	153.7	32.4	0.353	232.21	2.38	6569	147.2	0.50	49.8
1994	3	5.21	2.91	26.72	252	54.7	56.8	0.170	151.83	1.99	5895	77.6	0.48	30.9
1994	4	5.09	2.60	15.71	218	72.5	46.3	0.205	97.83	1.09	12783	54.5	0.49	17.4
1994	5	57.93	25.49	239.35	2261	102.3	25.0	2.073	1330.98	1.66	5821	803.5	0.53	247.9
1994	6	10.43	4.38	18.46	190	64.2	24.0	0.296	98.86	10.92	3137	74.2	0.49	18.8
1994	7	0.27	0.15	0.94	11	55.6	34.9	0.007	6.22	1.55	15793	3.7	0.55	1.1
1994	8	3.57	1.39	8.51	75	52.7	39.0	0.121	54.51	2.35	3963	30.7	0.57	10.1
1994	9	57.12	25.13	175.20	1854	102.0	23.4	1.792	1016.26	4.03	4054	629.2	0.45	189.1
1994	10	15.38	7.23	64.57	564	111.1	47.5	0.503	337.51	2.80	5286	203.0	0.46	81.8
1994	11	4.34	2.21	27.47	282	58.2	49.9	0.173	122.93	1.40	23476	61.4	0.55	31.2
1994	12	9.90	4.95	21.02	164	87.4	24.4	0.301	88.13	4.92	2722	58.2	0.52	19.4
1994	13	39.14	16.44	97.15	885	107.1	17.5	1.527	483.82	4.72	3499	304.0	0.47	90.4
1994	14	8.78	4.74	28.05	365	132.9	75.9	0.246	198.43	2.20	13948	108.2	0.56	26.9
1994	15	6.99	3.71	55.16	496	181.4	46.3	0.148	261.36	0.84	6831	156.5	0.46	72.6
1994	16	58.40	29.20	159.10	1520	118.9	41.3	1.813	1019.90	2.48	4868	650.1	0.54	139.5
1995	1	8.05	3.78	55.33	393	128.6	37.6	0.239	230.99	2.25	5800	130.1	0.48	52.6
1995	2	10.14	4.16	49.62	375	152.7	17.5	0.358	273.68	1.47	6752	172.0	0.50	60.8
1995	3	5.23	2.93	35.51	275	53.8	49.6	0.175	180.93	2.08	5928	91.8	0.48	39.6
1995	4	5.11	2.61	20.72	228	67.9	48.9	0.214	125.92	0.99	12785	68.3	0.50	26.4
1995	5	58.14	26.16	280.20	2428	103.1	26.0	2.092	1535.09	1.78	5892	922.8	0.53	301.2
1995	6	10.46	4.39	21.68	202	59.4	21.0	0.329	116.05	8.94	3259	86.9	0.49	20.7
1995	7	0.27	0.15	1.05	12	52.8	33.4	0.007	6.98	1.65	16011	4.2	0.55	1.1
1995	8	3.60	1.44	11.44	83	79.7	36.3	0.128	65.62	2.51	4120	35.0	0.55	13.4
1995	9	57.20	25.17	198.41	1960	96.4	21.9	1.775	1087.98	5.24	4165	668.8	0.45	223.7
1995	10	15.46	7.27	76.56	612	117.1	47.4	0.492	398.37	1.92	5374	238.0	0.46	96.9
1995	11	4.36	2.22	34.72	302	57.8	48.1	0.180	146.60	2.46	23800	72.4	0.55	39.6
1995	12	9.93	4.96	25.47	182	90.0	24.5	0.320	104.68	4.12	2902	67.9	0.52	24.5
1995	13	39.21	16.86	118.25	959	106.8	16.8	1.592	559.63	4.68	3594	347.4	0.46	118.4
1995	14	8.83	4.77	35.68	383	122.5	75.6	0.261	231.30	2.53	14106	121.1	0.57	38.3
1995	15	7.04	3.73	65.97	537	183.4	45.4	0.148	307.40	1.81	6916	185.1	0.47	87.5
1995	16	58.61	29.30	179.09	1623	125.5	40.7	1.821	1107.04	3.41	5016	699.9	0.54	158.3

Задание 1. Допустим, что в каждый момент времени t производство в каждой из стран задается производственной функцией Кобба-Дугласа с постоянной отдачей на масштаб:

$$Y(t) = K(t)^{\alpha} H(t)^{\beta} L(t)^{1-\alpha-\beta},$$

где Y - выпуск, K и H - объем физического и человеческого капитала, L - труд.

Эконометрическая модель может быть сформулирована либо для исходных показателей, либо в расчете на душу населения, либо в темповой записи.

Так, после логарифмирования производственной функции получим следующую эконометрическую модель:

$$\ln Y_{it} = \gamma_0 + \gamma_1 \ln K_{it} + \gamma_2 \ln H_{it} + \gamma_3 \ln L_{it} + \varepsilon_{it}. \qquad (1)$$

Обозначим:

$y_{it} = \frac{Y_{it}}{L_{it}}$ - ВВП на одного занятого в экономике;

$k_{it} = \frac{K_{it}}{L_{it}}$ - физический капитал (объем основных фондов на одного занятого);

$h_{it} = \frac{H_{it}}{L_{it}}$ - человеческий капитал в расчете на одного занятого;

$l_{it} = \frac{L_{i,t}}{L_{i,t-1}}$ - темп роста занятости в экономике.

Тогда модель можно записать в виде:

$$\ln y_{it} = \delta_0 + \delta_1 \ln k_{it} + \delta_2 \ln h_{it} + \delta_3 \ln l_{it} + \eta_{it}. \qquad (2)$$

Наконец, можно рассмотреть уравнение в темпах роста:

$$\ln \frac{y_{it}}{y_{it-1}} = \delta_0 + \delta_1 \ln \frac{k_{it}}{k_{it-1}} + \delta_2 \ln \frac{h_{it}}{h_{it-1}} + \delta_3 \ln \frac{l_{it}}{l_{it-1}} + \xi_{it}. \qquad (3)$$

Проверьте гипотезу, что $\delta_1 + \delta_2 = \delta_3$.

Каждую из моделей (1)-(3) можно расширить за счет квадратов переменных, перекрестных произведений и т.д.

Каковы достоинства и недостатки каждой из формулировок? Насколько хорошо предложенные переменные измеряют соответствующие факторы? Подумайте, в каком случае индивидуальные и временные эффекты, скорее всего, будут значимы, когда незначимы?

В зависимости от номера варианта (см. табл. 4) оцените коэффициенты модели (1), (2) или (3).

Таблица 4

Варианты для самостоятельной работы

Номер варианта	Номер модели	Независимые переменные	Номер варианта	Номер модели	Независимые переменные
1	1	x_2, x_4, x_6	16	2	x_2, x^2_2, x^2_4
2	1	x_2, x_4, x_7	17	2	x_2, x^2_2, x_4, x_6
3	1	x_1, x_4, x_6	18	2	x_2, x^2_2, x_4, x_7
4	1	x_1, x_4, x_7	19	2	x_2, x^2_2, x_4, x_6
5	1	x_2, x_4, x^2_4	20	2	x_2, x^2_2, x_4, x_7
6	1	x_2, x^2_2, x^2_4	21	3	x_2, x_3, x_6
7	1	x_2, x^2_2, x_4, x_6	22	3	x_2, x_3, x_7
8	1	x_2, x^2_2, x_4, x_7	23	3	x_1, x_3, x_6
9	1	x_2, x_4, x^2_4, x_6	24	3	x_1, x_3, x_7
10	1	x_2, x_2, x^2_4, x_7	25	3	x_2, x_3, x^2_3
11	2	x_2, x_4, x_6	26	3	x_2, x^2_2, x^2_3
12	2	x_2, x_4, x_7	27	3	x_2, x^2_2, x_3, x_6
13	2	x_1, x_4, x_6	28	3	x_2, x^2_2, x_3, x_7
14	2	x_1, x_4, x_7	29	3	x_2, x_3, x^2_3, x_6
15	2	x_2, x_4, x^2_4	30	3	x_2, x_2, x^2_3, x_7

Для каждого из вариантов выполните следующие этапы:

1. Выполните необходимые преобразования. Рассчитайте описательные статистики для каждого из показателей. Постройте диаграммы рассеивания. Насколько существенны различия между странами?
2. Найдите оценки коэффициентов модели методом наименьших квадратов;
3. Выполните within – трансформацию, рассчитайте коэффициенты модели с фиксированными эффектами, найдите исправленные стандартные ошибки и доверительные интервалы для коэффициентов, с помощью F – теста проверьте гипотезу о равенстве индивидуальных эффектов и о значимости факторов;
4. Выполните between – трансформацию, рассчитайте оценки коэффициентов уравнения between – регрессии;
5. Используя результаты within и between – регрессии, найдите оценку параметра θ. Оцените коэффициенты модели со случайными эффектами;
6. Проверьте гипотезу о значимости случайных эффектов с помощью LM – теста Бреуша-Пагана;
7. Выполните тест Хаусмана для сравнения оценок коэффициентов моделей с фиксированными и случайными эффектами.
8. Добавьте в модель временные эффекты, проверьте их значимость.
9. Обоснуйте выбор окончательной модели.

Задание 2. Оцените коэффициенты модели:

$$c_{it} = \alpha + \beta_1 y_{it} + \beta_2 p_{it} + \varepsilon_{it},$$

где c_{it} - логарифм потребления на душу населения;

y_{it} - логарифм ВВП на душу населения;

p_{it} - темп инфляции.

Задание 3. Постройте модель уровня потребления электроэнергии:

$$El_{it} = \alpha + \beta_1 y_{it} + \beta_2 p_{it} + \varepsilon_{it};$$

где El_{it} - логарифм потребления электроэнергии на душу населения;

y_{it} - логарифм ВВП на душу населения;

p_{it} - индекс потребительских цен.

Задание 4. Постройте модель зависимости нормы национальных сбережений (отношения сбережений к ВВП) от темпов роста ВВП, индекса потребительских цен, уровня демографической нагрузки. Подумайте, какие еще показатели следует включить в модель.

Список литературы

1. Айвазян С.А., Мхитарян В. С. Прикладная статистика и основы эконометрики: Учеб. для студентов экон. спец. вузов / С. А. Айвазян, В. С. Мхитарян; Гос. ун-т, Высш. шк. экономики. - М.: ЮНИТИ, 1998. - 1022 с.
2. Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы: Учебник. –М.: Финансы и статистика, 1998. –352с.
3. Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс. Учебник. –М.: Дело, 2000. – 400 с.
4. Baltagi B.H. Econometric Analysis of Panel Data. John Willey and Sons, New York., 1995.
5. Green W.H. Econometric Analysis. Third Edition. New York University, Pritice Hall, 1995. - 1070 p.
6. Matyas L. and Sevestre P. The econometrics of panel data. A Handbook of theory and application. 2nd revised edition, Kluwer Academic Publisher, Dordrecht.
7. Verbeek M. A guide of modern econometrics. John Willey and Sons, New York, 2000. – 386 p.

Об авторах:

Балаш Владимир Алексеевич – доктор экономических наук, профессор кафедры банковского дела Саратовского государственного социально-экономического университета, тьютор Московского государственного университета экономики, статистики и информатики.

Балаш Ольга Сергеевна – кандидат экономических наук, доцент, заведующая кафедрой высшей математики и информационных технологий Саратовского института (филиала) Московского государственного университета коммерции, тьютор Московского государственного университета экономики, статистики и информатики.