Предварительный анализ данных

Данные (1)

PSID – Panel Study of Income Dynamics
(University of Michigan)

Список исходных переменных:

- ***pid*** идентификатор индивида
- ***wave*** идентификатор волны
- ***logpay*** логарифм заработной платы человека за месяц до опроса
- ***pnjuwks*** число недель, проведенных без работы в предыдущем году
- ***sex*** пол
- ***age*** возраст
- ***agesq*** возраст в квадрате

- **`xtset pid wave`**
- `panel variable: pid (unbalanced)`
- `time variable: wave, 1 to 11, but with gaps`
- `delta: 1 unit`

Описание структуры панели (1)

- **xtdes**

```
     pid:  10002251, 10004491, ..., 1.194e+08        n =  27400
    wave:  1, 2, ..., 11                             T =     11
           Delta(wave) = 1 unit
           Span(wave)  = 11 periods
           (pid*wave uniquely identifies each observation)

Distribution of T_i:   min    5%   25%     50%    75%    95%   max
                         1     1     9      10     11     11    11

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+-------------
    12900     47.08   47.08 |  11111111111
     7063     25.78   72.86 |  111111111..
     4275     15.60   88.46 |  ..........1
     1014      3.70   92.16 |  .........11
     1012      3.69   95.85 |  1111111111.
      325      1.19   97.04 |  111111111.1
      294      1.07   98.11 |  .........1.
      197      0.72   98.83 |  ........111
      140      0.51   99.34 |  ......11111
      180      0.66  100.00 | (other patterns)
 ---------------------------+-------------
    27400    100.00         |  XXXXXXXXXXX
```

Описательные статистики (1)

- **xtsum**

```
Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
pid      overall | 4.66e+07   3.73e+07   1.00e+07   1.19e+08 |     N =  227107
         between |            4.24e+07   1.00e+07   1.19e+08 |     n =   27400
         within  |                  0   4.66e+07   4.66e+07 | T-bar = 8.2885
wave     overall | 5.855244   3.106436          1         11 |     N =  227107
         between |            2.168274          5         11 |     n =   27400
         within  |            2.936897   .8552444   11.25524 | T-bar = 8.28858
logpay   overall | 6.585763   .8622559  -.3439421   10.97223 |     N =   65692
         between |            .8510987  -.1067718   9.214394 |     n =   16056
         within  |            .3713454   2.267388   9.820659 | T-bar = 4.09143
sex      overall |  1.53566   .4987286          1          2 |     N =  130075
         between |            .4992113          1          2 |     n =   27400
         within  |            .0022639   .8689935   1.868994 | T-bar = 4.74726
age      overall | 44.61089    18.5723         -9        101 |     N =  130075
         between |            19.27007         -9         97 |     n =   27400
         within  |            2.553033  -4.055773   69.94423 | T-bar = 4.74726
pnjuwks  overall | 1.638432   8.644212         -9   52.28571 |     N =  102675
         between |            7.302418         -9   52.28571 |     n =   19657
         within  |            6.137296  -50.70442     50.667 | T-bar = 5.22333
```

Данные (2)

- Citydata - данные о ценах на продукты и некоторые другие товары и услуги, собираемые корреспондентами журнала “Economist” в столицах и крупнейших городах мира, дополненные сведениями о ВВП и ВВП по ППС

- **`xtset country t`**
- `panel variable: country (strongly balanced)`
- `time variable: t, 1993 to 2008`
- `delta: 1 unit`

Описание структуры панели (2)

- **xtdes**

```
 country:  1, 2, ..., 31                     n =       31
       t:  1993, 1994, ..., 2008              T =       16
           Delta(t) = 1 unit
           Span(t)  = 16 periods
           (country*t uniquely identifies each observation)

Distribution of T_i:   min    5%   25%   50%   75%   95%  max
                        16    16    16    16    16    16   16

      Freq.  Percent    Cum. |  Pattern
 ----------------------------+------------------
        31    100.00  100.00 |  1111111111111111
 ----------------------------+------------------
        31    100.00         |  XXXXXXXXXXXXXXXX
```

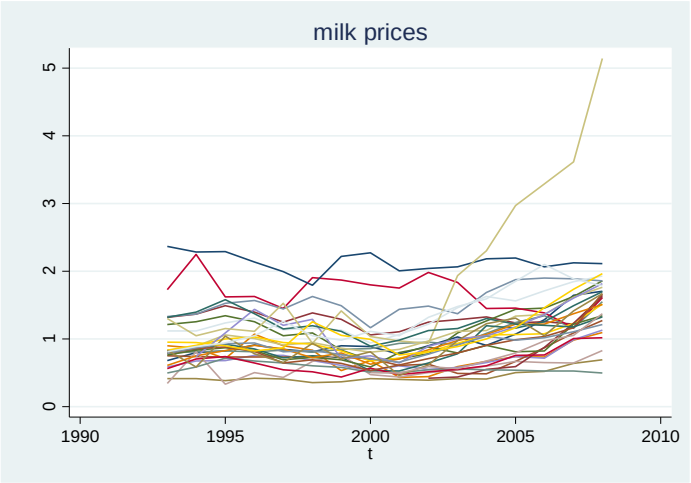
Описательные статистики (2)

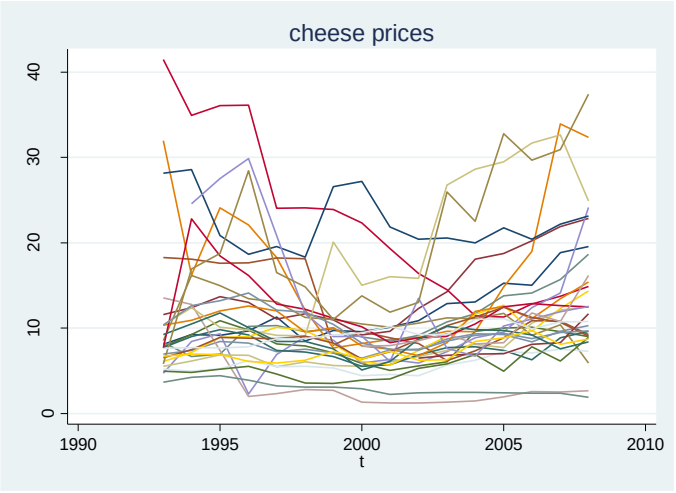
- **xtsum country t gdp gdp_ppc milk cheese**

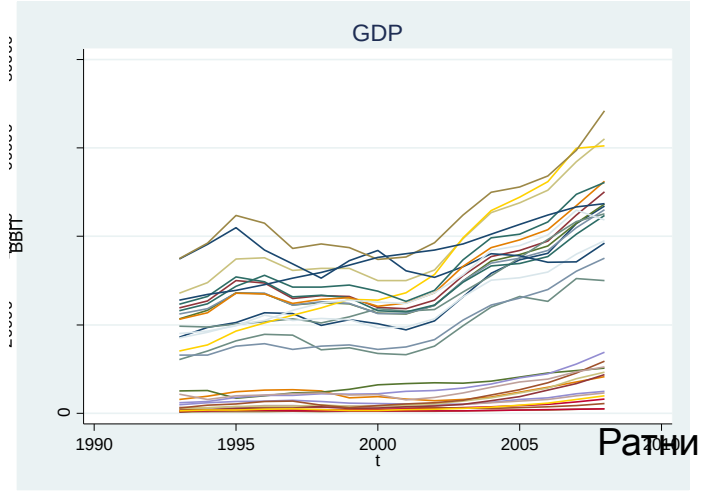
```
Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
country  overall |        16   8.953302          1         31 |     N =      496
         between |             9.092121          1         31 |     n =       31
         within  |                   0         16         16 |     T =       16
                 |                                            |
t        overall |    2000.5   4.614426       1993       2008 |     N =      496
         between |                   0     2000.5     2000.5 |     n =       31
         within  |             4.614426       1993       2008 |     T =       16
                 |                                            |
gdp      overall |  18060.56   15814.53    298.126   68433.13 |     N =      496
         between |              14709.7   530.6633   44412.66 |     n =       31
         within  |             6346.723  -1058.355   45338.77 |     T =       16
                 |                                            |
gdp_ppc  overall |  18084.51    11899.9    949.582   47439.93 |     N =      496
         between |             11268.33   1623.989   35881.99 |     n =       31
         within  |             4298.847     3652.6   32244.25 |     T =       16
                 |                                            |
milk     overall |   1.05488   .5091208   .3333333   5.143306 |     N =      462
         between |             .3910025   .4470736   2.134499 |     n =       30
         within  |              .327886   .0688604   4.279507 | T-bar =     15.4
                 |                                            |
cheese   overall |  11.05028   6.656141       1.25   41.46653 |     N =      461
         between |             5.053736   2.964437   22.39296 |     n =       30
         within  |             4.347618  -3.763969   30.17193 | T-bar = 15.3667
.
```

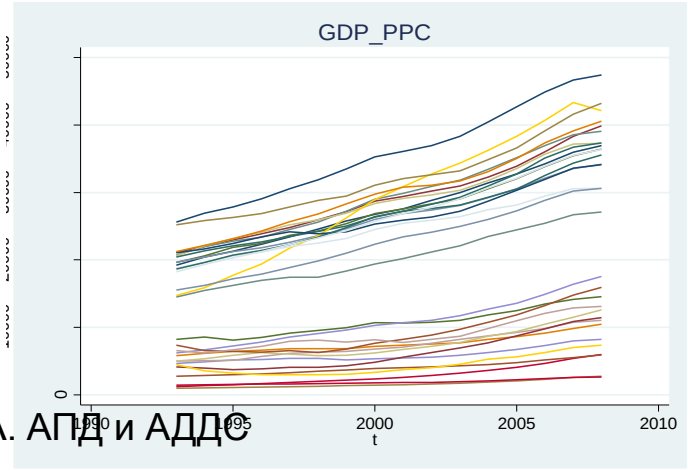
Визуальный анализ (2)

- `xtline milk, overlay legend(off) title(milk prices)`
- `xtline gdp, overlay legend(off) title(GDP)`



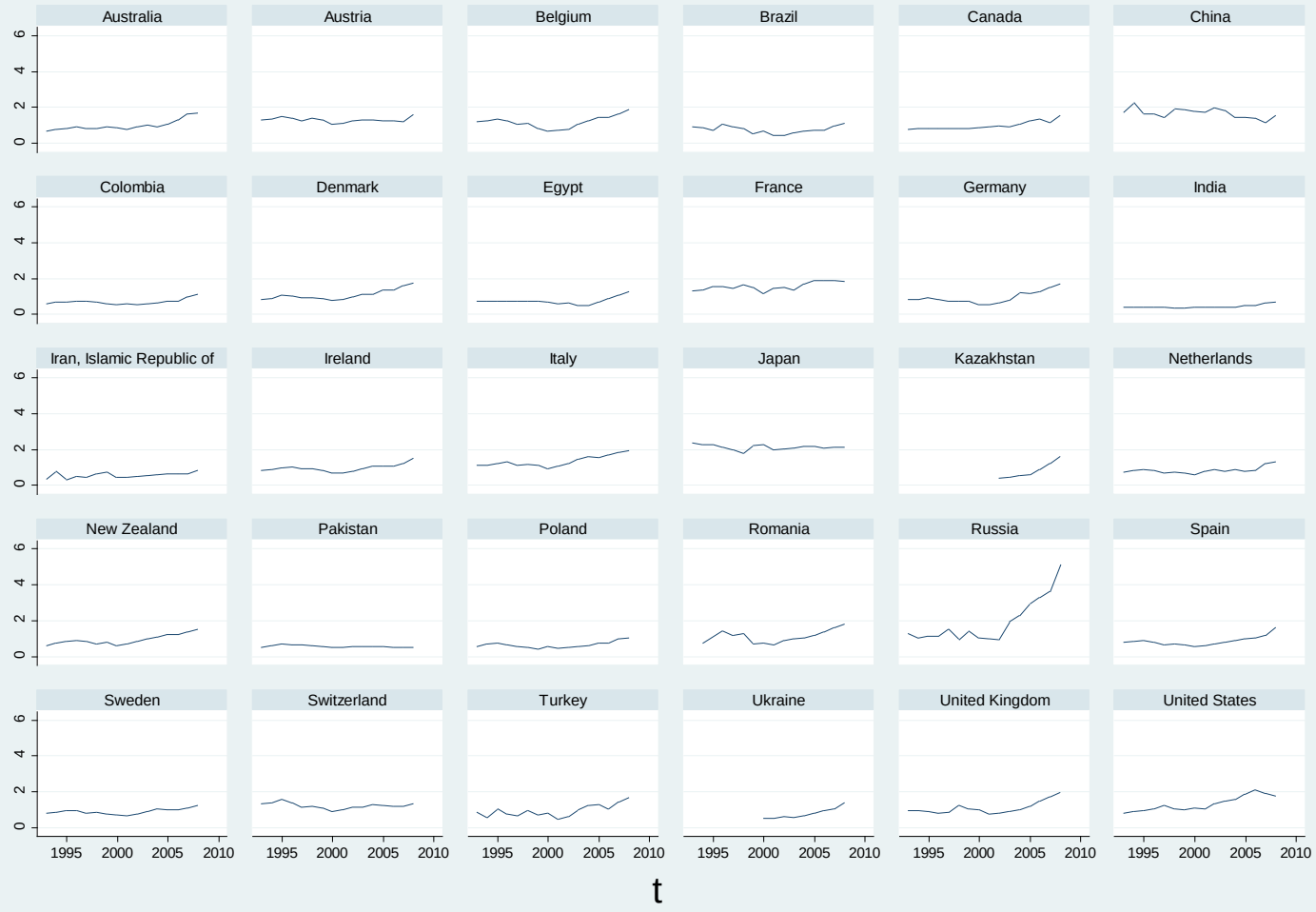






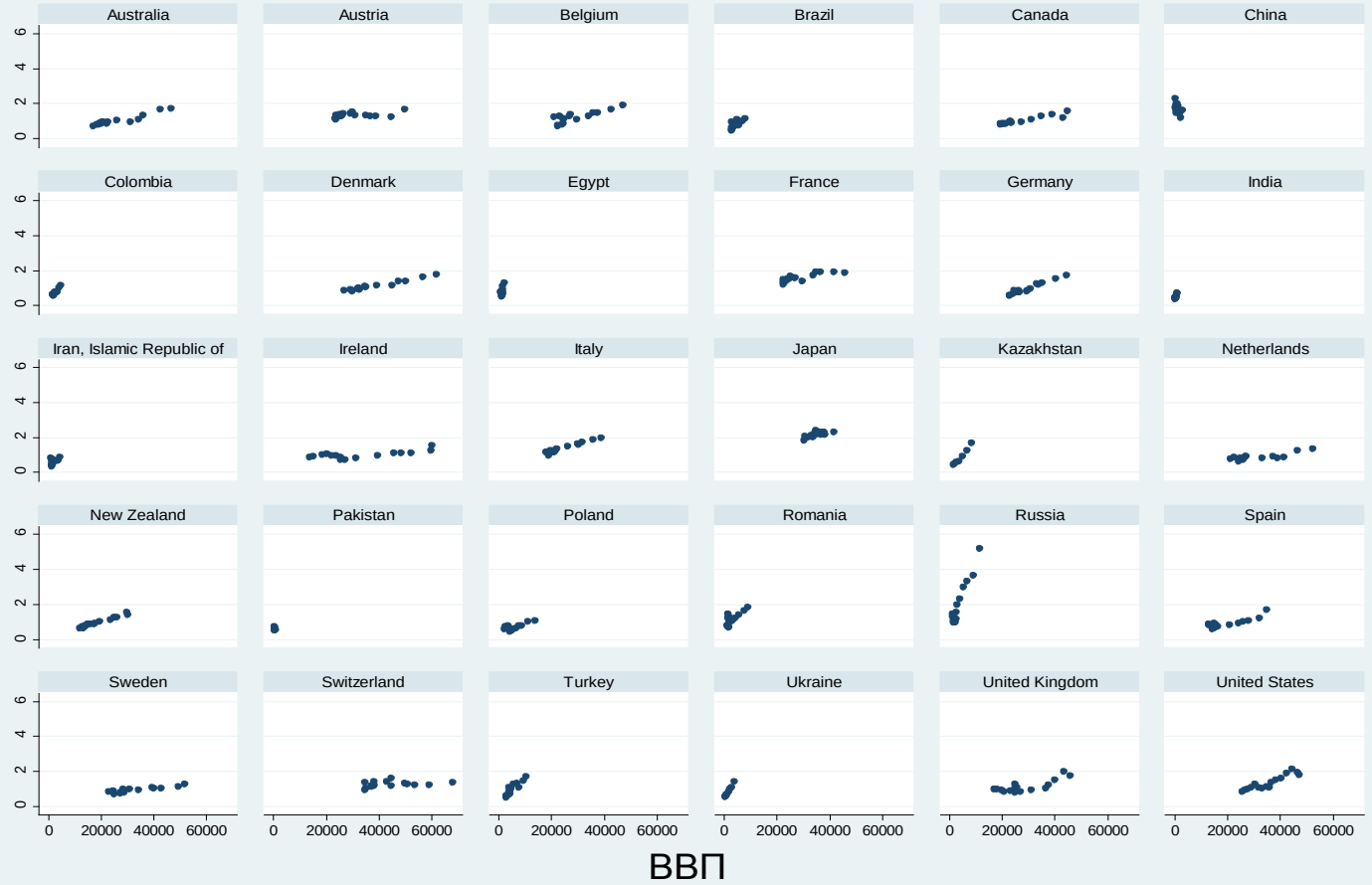
Визуальный анализ (2)

- `xtline milk, i(name) t(t)`



Визуальный анализ (2)

- `xtline milk, recast(scatter) i(name) t(gdp)`



Тестирование гомогенности коэффициентов

(анализ возможности объединения данных в панель)

Постановка задачи: проверка возможности объединения данных в панель по объектам

Тестирование соответствия данных
одной из трех гипотетических спецификаций:

- модель без ограничений (0) $y_{it} = X_{it}\beta_i + \alpha_i + u_{it}$
  (регрессия с гетерогенными по объектам коэффициентами наклона и свободным членом),
- модель с ограничениями (1) $y_{it} = X_{it}\beta + \alpha_i + u_{it}$
  (регрессия с детерминированным индивидуальным эффектом),
- модель с ограничениями (2) $y_{it} = X_{it}\beta + \alpha + u_{it}$
  (сквозная регрессия).

Оценки параметров модели (0) без ограничений

- Пусть $y_{i\bullet} = \frac{1}{T}\sum_{t=1}^{T} y_{it}, \quad x_{i\bullet} = \frac{1}{T}\sum_{t=1}^{T} x_{it}$
- Оценки МНК β_i и α_i

$$\begin{cases} \hat{\beta}_i = W_{xx,i}^{-1} W_{xy,i} \\ \hat{\alpha}_i = y_{i\bullet} - \hat{\beta}_i' x_{i\bullet} \\ i = \overline{1,N} \end{cases} \quad где$$

$$W_{xx,i} = \sum_{t=1}^{T} (x_{it} - x_{i\bullet})(x_{it} - x_{i\bullet})' = x_i'(I_T - \tfrac{J_T}{T})x_i$$

$$W_{xy,i} = \sum_{t=1}^{T} (x_{it} - x_{i\bullet})(y_{it} - y_{i\bullet})' = x_i'(I_T - \tfrac{J_T}{T})y_i$$

$$W_{yy,i} = \sum_{t=1}^{T} (y_{it} - y_{i\bullet})^2 = y_i'(I_T - \tfrac{J_T}{T})y_i$$

- Это оценки группы «within».
- Сумма квадратов остатков модели без ограничений:

$$S_0 = \sum_{i=1}^{N} RSS_i, \quad где \quad RSS_i = W_{yy,i} - W_{xy,i}' W_{xx,i}^{-1} W_{xy,i}$$

Оценки параметров модели с ограничением (1)

- Оценки МНК регрессии (1) – это оценки FE

$$\begin{cases} \hat{\beta}_W = W_{xx}^{-1} W_{xy} \\ \hat{\alpha}_i = y_{i\bullet} - \hat{\beta}_W^{'} x_{i\bullet} \\ i = \overline{1, N} \end{cases} \quad \text{где} \quad \begin{aligned} W_{xx} &= \sum_{i=1}^{N} W_{xx,i} = x'Wx \\ W_{xy,i} &= \sum_{i=1}^{N} W_{xy,i} = x'Wy \\ W_{yy,i} &= \sum_{i=1}^{N} W_{yy,i} = y'Wy \end{aligned}$$

- Сумма квадратов остатков модели с ограничениями:

$$S_1 = W_{yy} - W_{xy}^{'} W_{xx}^{-1} W_{xy}$$

Оценки параметров модели с ограничением (2)

- Это МНК оценки обыкновенной (сквозной) регрессии в отклонениях от глобального среднего

$$\begin{cases} \hat{\beta} = T_{xx}^{-1} T_{xy} \\ \hat{\alpha} = y_{\bullet\bullet} - \beta' x_{\bullet\bullet} \end{cases} \quad где \quad T_{xx} = \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - x_{\bullet\bullet})(x_{it} - x_{\bullet\bullet})' = x' T^* x$$

$$T_{xy} = \sum_{i=1}^{N} \sum_{t=1}^{T} (x_{it} - x_{\bullet\bullet})(y_{it} - y_{\bullet\bullet})' = x' T^* y \ ,$$

$$T_{yy} = \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - y_{\bullet\bullet})^2 = y' T^* y$$

$$y_{\bullet\bullet} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} y_{it} \qquad x_{\bullet\bullet} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it}$$

- Сумма квадратов остатков модели с ограничением (2):

$$S_2 = T_{yy} - T_{xy}^{'} T_{xx}^{-1} T_{xy}$$

Тестовые статистики

- Для проверки ограничения (1) используется F-тест:

.

$$F_1 = \frac{(S_1 - S_0)/[(N-1)K]}{S_0/[NT - N(K+1)]} \overset{(1)}{\sim} F(\,(N-1)K,\, NT - N(K+1)\,)$$

- Для проверки ограничения (2) используется F-тест:

.

$$F_2 = \frac{(S_2 - S_0)/[(N-1)(K+1)]}{S_0/[NT - N(K+1)]} \overset{(2)}{\sim} F(\,(N-1)(K+1),\, NT - N(K+1)\,)$$

Тестовые статистики

- Логика исследования:

  если гипотеза (2) отвергается, то проверяется гипотеза (1),

  если (1) отвергается, нужно оценивать регрессию без ограничений,

  если же гипотезу (1) нет оснований отвергнуть, то

  проверяется гипотеза о гомогенности свободного члена при условии, что гипотеза о гомогенности наклона выполнена

- H0: $\alpha_1 = \alpha_2 = ... = \alpha_N \quad | \quad \beta_1 = \beta_2 = ... = \beta_N$

- Для проверки ограничения H4 используется F-тест:

$$F_3 = \frac{(S_2 - S_1)/(N-1)}{S_1/[N(T-1)-K]} \overset{H0}{\sim} F(N-1, N(T-1)-K)$$

- Аналогично можно исследовать модель, где коэффициенты ведут себя одинаково для всех объектов, но изменяются со временем.

Постановка задачи: проверка возможности объединения данных в панель по времени

Тестирование соответствия данных
одной из трех гипотетических спецификаций:

- модель без ограничений (0) $y_{it} = X_{it}\beta_t + \alpha_t + u_{it}$
  (регрессия с гетерогенными по времени коэффициентами наклона и свободным членом),
- модель с ограничениями (1) $y_{it} = X_{it}\beta + \alpha_t + u_{it}$
  (регрессия с детерминированным временным эффектом),
- модель с ограничениями (2) $y_{it} = X_{it}\beta + \alpha + u_{it}$
  (сквозная регрессия).

Данные (1)

PSID – Panel Study of Income Dynamics
(University of Michigan)

Список исходных переменных:

- ***pid*** идентификатор индивида
- ***wave*** идентификатор волны
- ***logpay*** логарифм заработной платы человека за месяц до опроса
- ***pnjuwks*** число недель, проведенных без работы в предыдущем году
- ***sex*** пол
- ***age*** возраст
- ***agesq*** возраст в квадрате

- **`xtset pid wave`**
- `panel variable: pid (unbalanced)`
- `time variable: wave, 1 to 11, but with gaps`
- `delta: 1 unit`

Код STATA
для тестирования возможности объединения данных в панель (1)

```
/* усреднение по индивидам в каждой волне */
egen mtX=mean(X), by(wave)
/* усреднение по времени для каждого индивида */
egen miX=mean(X), by(pid)
/* вычисление отклонений от средних */
gen diX=X-miX
gen dtX=X-mtX
/* оценивание модели (0) без ограничений */
reg dtlogpay dtpnjuwks dtage dtagesq if wave==2
scalar z2=e(rss)
reg dtlogpay dtpnjuwks dtage dtagesq if wave==3
scalar z3=e(rss)
reg dtlogpay dtpnjuwks dtage dtagesq if wave==4
scalar z4=e(rss)
reg dtlogpay dtpnjuwks dtage dtagesq if wave==5
scalar z5=e(rss)
scalar tot=z2+z3+z4+z5
```

Код STATA
для тестирования возможности объединения данных в панель (2)

```
/* оценивание модели с ограничением (1) */
regr dtlogpay dtpnjuwks dtage dtagesq
scalar z6= e(rss)
/* оценивание модели с ограничением (2) */
regr logpay pnjuwks age agesq
scalar z7 = e(rss)
/* вычисление тестовых статистик и их p-values */
scalar ddf = 1324*4-16
scalar fh1=((z6-tot)/(9))/(tot/ddf)
scalar pval1 = Ftail(9,ddf,fh1)
scalar fh2 =((z7-tot)/(12))/(tot/ddf)
scalar pval2 = Ftail(12,ddf,fh2)
scalar fh3 =((z7-z6)/(3))/(z6/(ddf+9))
scalar pval3 = Ftail(3,ddf+9,fh3)
/* просмотр результатов */
scalar list pval1 pval2 pval3 ddf fh1 fh2 fh3
```

Интерпретация результатов

- fh1 = 0.81121957
- pval1 = 0.60582507

- Статистика F1 сопоставляет модель без ограничений (0) и модель с детерминированным временным эффектом (1).
- Её p-value показывает, что вероятность ошибиться, отвергнув гипотезу об эквивалентности моделей (0) и (1) равна примерно 60%.
- Вывод: нет оснований отвергать гипотезу (1), и следует отдать предпочтение модели с детерминированным временным эффектом (1).

Интерпретация результатов

- fh2 = 3.409128
- pval2 = .00005185

- Статистика F2 сопоставляет модель без ограничений (0) и модель с гомогенными коэффициентами (2).
- Её p-value показывает, что вероятность ошибиться, отвергнув гипотезу об эквивалентности моделей (0) и (2) равна примерно 0%.
- Вывод: следовательно есть все основания отвергнуть гипотезу (2) и отдать предпочтение модели без ограничений (0).

Интерпретация результатов

- fh3 = 11.20455
- pval3 = 2.448e-07

- Статистика F3 сопоставляет FE-модель (1) и модель с гомогенными коэффициентами (2).
- Её p-value показывает, что вероятность ошибиться, отвергнув гипотезу об эквивалентности моделей (1) и (2) равна 0%.
- Вывод: следовательно есть все основания отвергнуть гипотезу (2) и отдать предпочтение модели с детерминированным временным эффектом (1).

Окончательные выводы

- Вывод:

  данные объединимы в панель, но необходимо принимать во внимание временной эффект, т.е. учитывать временные структурные сдвиги заработной платы.