小白学统计|面板数据分析与Stata应用笔记（三）

长面板数据分析

上两篇笔记我们讲到了短面板数据分析。短面板数据分析主要关注对不可观测的个体效应的处理，而对于误差自相关、异方差和截面相关的问题只提供经过校正的标准误。

与短面板数据不同，长面板数据分析主要关注对误差项的处理（因为时间T大），而将个体效应用虚拟变量来控制（因为个体n小）。

所以，对于长面板数据分析，我们不需要在固定效应模型、随机效应模型和混合回归模型之间进行选择，长面板数据分析先验假定长面板数据模型就是固定效应模型。

此外，需要注意的是，短面板数据分析对于时间效应，用虚拟变量来控制，而长面板数据分析，由于时间T相对较长，为避免损失较多的自由度，所以一般则用时间趋势项来控制。

可以认为长面板数据模型是一个特殊的双向固定效应模型。在这个模型中，个体效应用虚拟变量控制，时间效应用时间趋势项控制，长面板数据模型的估计主要关注对误差项的处理。

一、长面板数据模型的估计方法

通常有三种方法对长面板数据模型进行估计。

第一种：使用OLS估计这个特殊的双向固定效应模型，并对误差项的自相关、异方差和截面相关的问题只提供面板校正的标准误（使用命令xtscc或xtpcse命令实现），这种估计方法最为稳健。

第二种：如果存在自相关、异方差和截面相关的问题，则使用FGLS估计这个特殊的双向固定效应模型，这种方法只是解决了误差项自相关的问题，而并未考虑异方差或截面相关的问题，对于误差项的异方差和截面相关的问题仍然只是提供面板校正的标准误（使用命令xtpcse实现），这种估计方法介于稳健和效率之间。

第三种：使用FGLS估计这个特殊的双向固定效应模型，对误差项的自相关、异方差和截面相关的问题一并加以处理（使用命令xtgls实现），这种估计方法最有效率。

二、长面板数据模型的Stata估计命令

常用的估计长面板数据模型的Stata命令有三个:【xtpcse】、【xtgls】和【xtscc】

对于【xtscc】命令，我们在前两篇短面板数据的笔记中已经讲过，【xtscc】也适用于长面板数据分析，它可以实现长面板数据模型的第一种估计方法，对误差项的自相关、异方差和截面相关问题提供面板校正的标准误。

下面，我们讲一下【xtpcse】和【xtgls】估计命令

1、【xtpcse】命令

基本命令格式：xtpcse depvar indepvars,options

#命令的关键在于选项（options）,不同的选项可以处理不同的问题。

对于误差项三大问题【xtpcse】命令选项（options）的使用

（1）自相关问题（一阶自相关）

a.使用选项：corr(ar1)，使用的估计方法为FGLS

#误差项存在自相关时使用该选项；当T不比n大很多时使用该选项，因为此时T可能无法提供足够多的信息去估计每个个体的自相关系数，所以约束了每个个体的自相关系数都相等

b.使用选项：corr(psar1)，使用的估计方法为FGLS。

#误差项存在自相关时使用该选项；当T比n大很多时使用该选择项，当T比n大很多时每个个体的自相关系数可以不同，就可以使用选项

c.使用选项：corr(independent)或corr(ind)，使用的估计方法为OLS。

#误差项不存在自相关时，使用该选项

（2）异方差与截面相关问题

a.使用选项：independent

#误差项不存在异方差和截面相关问题，使用该选项

b.使用选项：hetonly（提供考虑异方差的面板校正标准误）

#误差项存在异方差但不存在截面相关问题，则使用该选项

c.使用选项：不加选项即可（提供既考虑异方差又考虑截面相关的面板校正标准误）

#误差项存在异方差和截面相关问题时，不加任何选项

选项：corr(ind)+independent等价于LSDV

2、【xtgls】命令

基本命令格式：xtgls depvar indepvars,options

#如果对误差项的处理正确，那么【xtgls】比x【tpcse】估计效果更好

对于误差项三大问题【xtgls】命令选项（options）的使用

（1）自相关问题（一阶自相关）

【xtgls】与【xtpcse】命令的选项对自相关问题的处理是相同的

a.使用选项：corr(ar1)，使用的估计方法为FGLS

#误差项存在自相关时使用该选项；当T不比n大很多时使用该选项，因为此时T可能无法提供足够多的信息去估计每个个体的自相关系数，所以约束了每个个体的自相关系数都相等

b.使用选项：corr(psar1)，使用的估计方法为FGLS。

#误差项存在自相关时使用该选项；当T比n大很多时使用该选择项，当T比n大很多时每个个体的自相关系数可以不同，就可以使用选项

c.使用选项：corr(independent)或corr(ind)，使用的估计方法为OLS。

#误差项不存在自相关时，使用该选项

（2）异方差与截面相关问题

a.使用选项：panels(iid)

#误差项不存在异方差和截面相关，使用该选项

b.使用选项：panles(heteroskedastic)

#误差项存在异方差但不存在截面相关问题时，使用该选项

c.使用选项：panels(correlated)#只适用于长面板数据

#误差项存在异方差和截面相关问题时，使用该选项

选项：corr(ind)+panels(iid)等价于LSDV

三、长面板数据分析的实例操作

#以数据集“mus08cigar.dta”为例估计香烟需求函数，数据来源于慕课上浙江大学方红生教授的面板数据分析与Stata应用课程中。

“mus08cigar.dta”数据集包括了美国10个州1963-1992年有关香烟消费量的相关变量。

参考上一篇文章短面板数据分析的基本程序，我们对长面板数据进行分析。

第一步 模型设定与数据

长面板数据不需要进行模型的选择，我们构造一个双向固定效应模型

$$\ln c_{it} = \beta_0 + \beta_1 \ln p_{it} + \beta_2 \ln pmin_{it} + \beta_3 \ln y_{it} + \mu_i + \gamma_t + \varepsilon_{it}$$

其中，被解释变量lnc为人均香烟消费量的对数，解释变量：lnp为实际香烟价格的对数，lnpmin为相邻州最低香烟价格的对数，lny为人均可支配收入的对数。

在Stata软件中对数据进行分析，执行如下步骤：

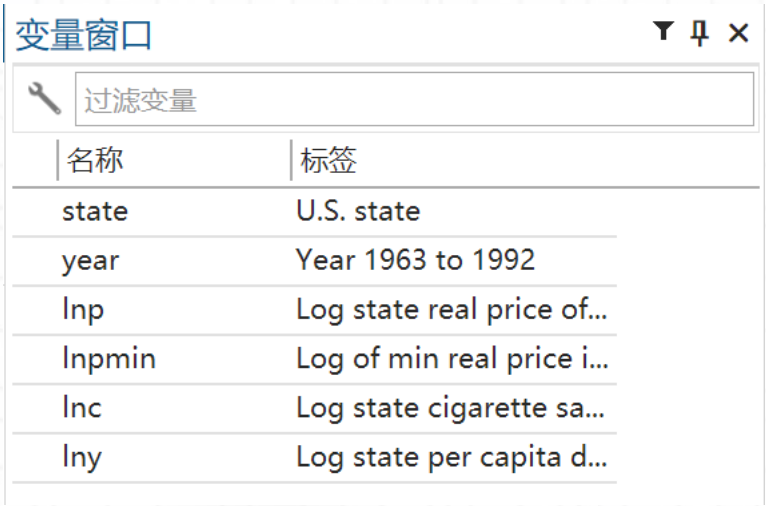
1、导入数据到Stata中

在Stata的“命令窗口”中输入

命令【use"数据集路径\mus08cigar.dta"】将“traffic.dta”数据集导入到Stata中，

例如【use"C:\Users\mus08cigar.dta"】。

将数据导入Stata后，即可在Stata的“变量窗口”中看到“mus08cigar”数据集中的各个变量的名称及其标签。



2、查看数据

在Stata的“命令窗口”输入命令【des】、【xtdes】查看“mus08cigar”数据集。

```
  obs:           300
 vars:             6                          26 Nov 2008 17:14
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
state           float   %9.0g                 U.S. state
year            float   %9.0g                 Year 1963 to 1992
lnp             float   %9.0g                 Log state real price of pack of cigarettes
lnpmin          float   %9.0g                 Log of min real price in adjoining states
lnc             float   %9.0g                 Log state cigarette sales in packs per capita
lny             float   %9.0g                 Log state per capita disposable income
-------------------------------------------------------------------------------
```

```
   state:  1, 2, ..., 10                                 n =         10
    year:  63, 64, ..., 92                               T =         30
           Delta(year) = 1 unit
           Span(year)  = 30 periods
           (state*year uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%      max
                        30      30      30        30        30      30       30

     Freq.  Percent     Cum. |  Pattern
----------------------------+------------------------------
        10   100.00  100.00 |  1111111111111111111111111111
----------------------------+------------------------------
        10   100.00         |  XXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

从输出结果我们可以看到：“mus08cigar”数据集包含300个观测值，6个变量。

面板数据的截面数 $n=10$,时间数 $T=30$, $T>n$,说明这是一个长面板数据集。

输入命令【xtset state year】,告诉Stata软件，这是一个以截面变量state为州，时间变量为year的面板数据。

```
. xtset state year
       panel variable:  state (strongly balanced)
        time variable:  year, 63 to 92
                delta:  1 unit
```

由“strongly balance”可知，这是一个平衡面板数据。

至此，我们可以知道，“mus08cigar”数据集是一个10个州，1963-1992年的长面板数据集且为平衡面板数据集。

第二步 描述性统计作图

1、描述性统计

使用命令【sum 关键变量】可以得到关键变量的描述性统计表。

在Stata中输入命令【sum lnc lnp lnpmin lny】,得到解释变量与被解释变量的观测值、均值、标准差、最小值和最大值。

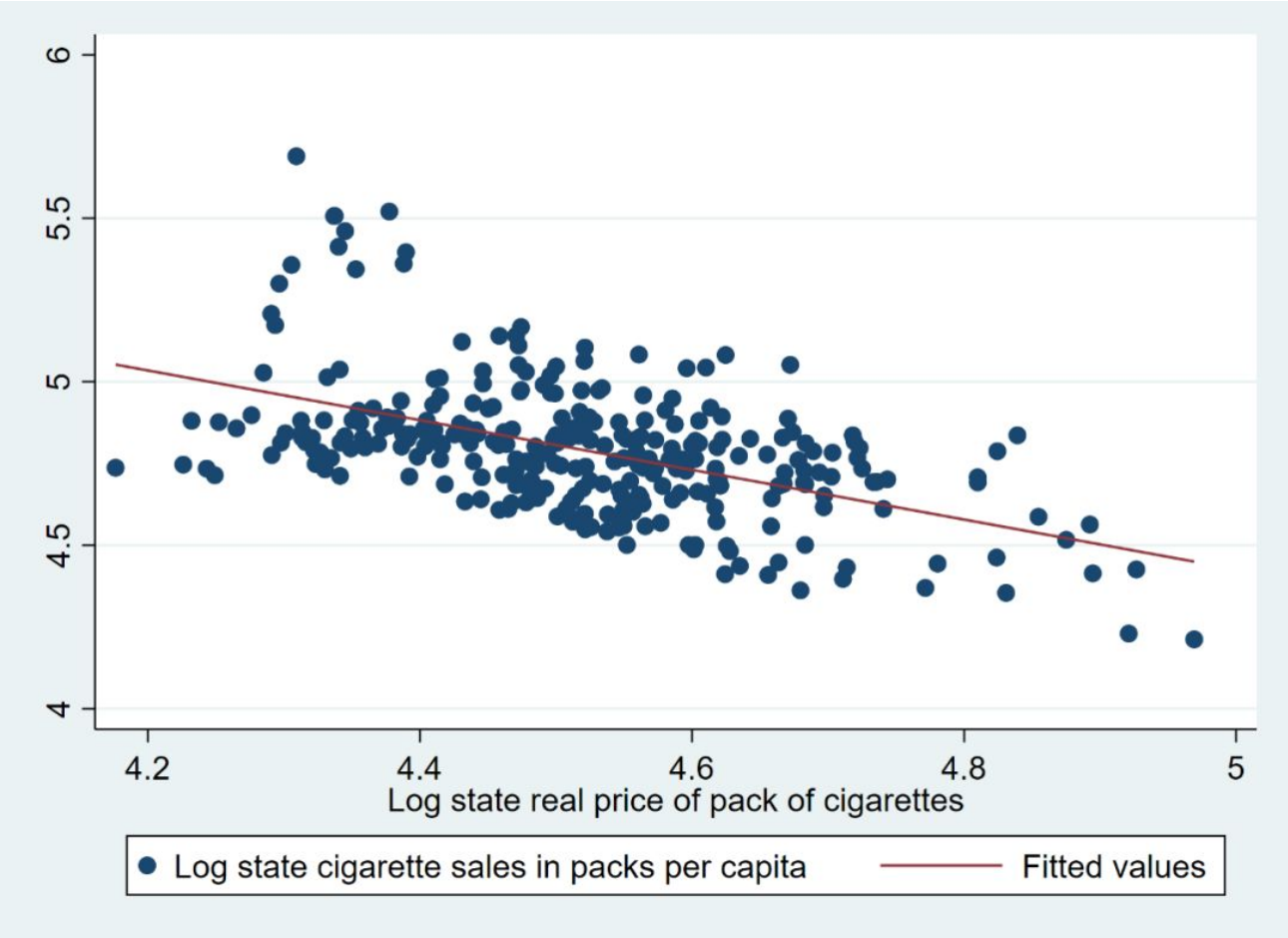
```
. sum lnc lnp lnpmin lny

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         lnc |        300    4.792591    .2071792    4.212128   5.690022
         lnp |        300    4.518424    .1406979    4.176332    4.96916
      lnpmin |        300      4.4308    .1379243      4.0428   4.831303
         lny |        300    8.731014    .6942426    7.300023    10.0385
```

2、绘制散点图及回归直线

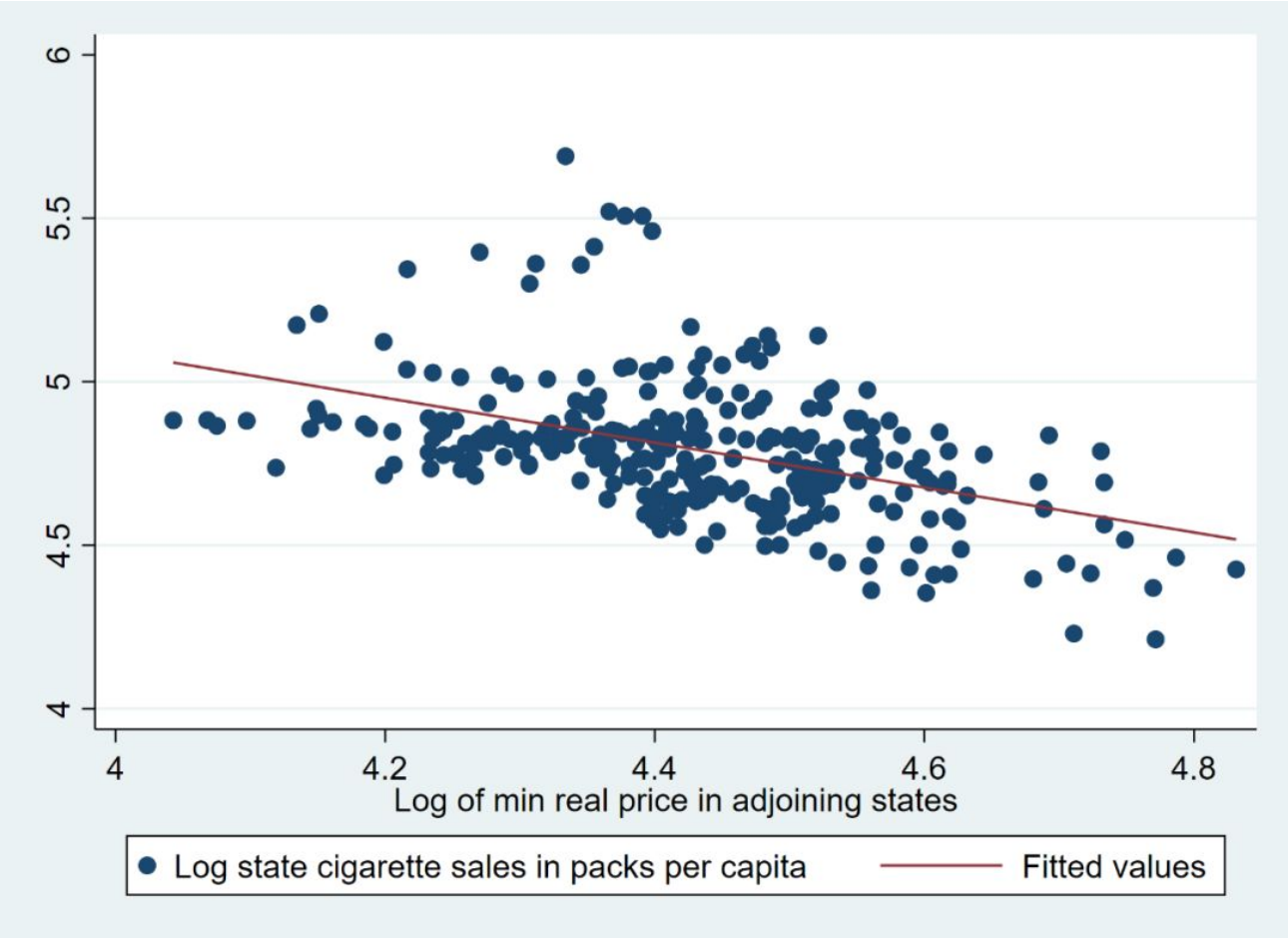
在回归之前，我们先画出核心变量lnp与被解释变量lnc的散点图及回归直线，来预先观测一下核心变量与被解释变量之间是否存在理论上预期的负相关系。

使用命令【twoway(scatter lnc lnp)(lfit lnc lnp)】画出核心变量“lnp”与被解释变量“lnc”的散点图及回归直线。



由结果可知，lnp与lnc之间是负相关系的，与理论预期一致。

接下来，我们做出相邻州的香烟价格的对数与被解释变量的散点图及回归直线，看一下核心变量lnpmin与被解释变量lnc之间是否存在理论上预期的正相关系。使用命令【twoway(scatter lnc lnpmin)(lfit lnc lnpmin)】

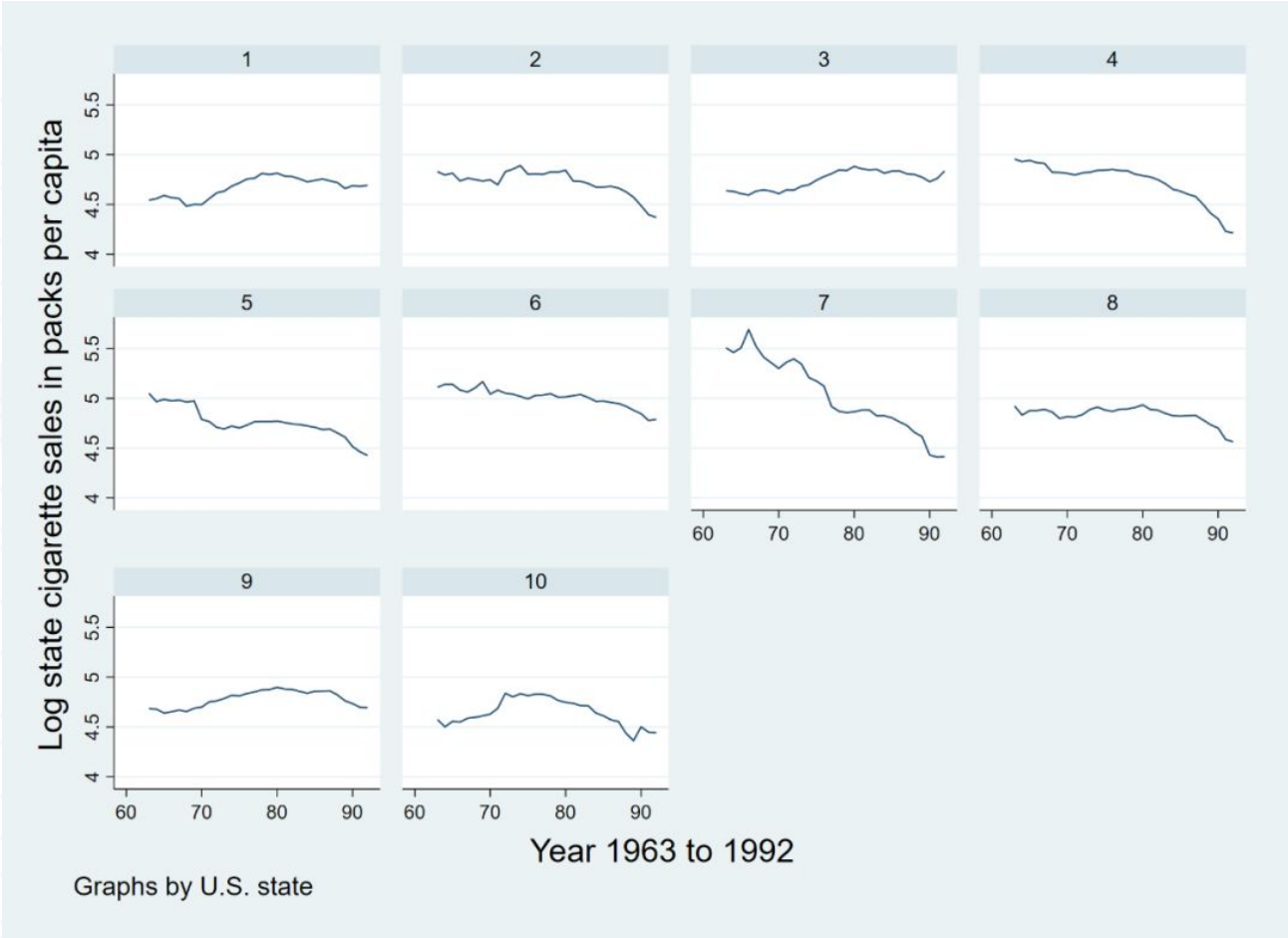


由结果可知，相邻州的香烟价格的对数lnpmin与被解释变量lnc之间是正相关系的，这与我们的理论预期并不符合。

不过，因为我们并没有控制其他的影响因素，所以这个结果并不是完全正确的，在之后的操作中，我们可以使用命令【avplot】绘制变量之间的偏相关图。

3、绘制核心变量的时间序列图

使用命令【xtline lnc】做出核心变量人均香烟消费的对数lnc在各个州的时间序列图，以研究分析人均香烟消费的对数lnc在每个州中的变动趋势。



观察lnc在各个州的时序图，我们可以发现，1980年之后，所有州的人均香烟消费率基本都呈现出下降趋势。

使用命令【xtline lnp】做出美国10个州1963-1992年实际香烟价格对数的时间序列图。

观察发现：1980年之后，所有州的香烟价格基本都呈现了上升的趋势。

第三步 模型估计

首先，我们先假定不存在自相关、异方差和截面相关这三大问题，使用LSDV估计双向固定效应模型。

依次进行如下操作：

使用命令【tab state,gen(state)】生成州虚拟变量；

使用命令【gen t=year-62】生成时间趋势变量；

输入命令【reg lnc lnp lnpmin lny state2-state10 t】进行LSDV估计；

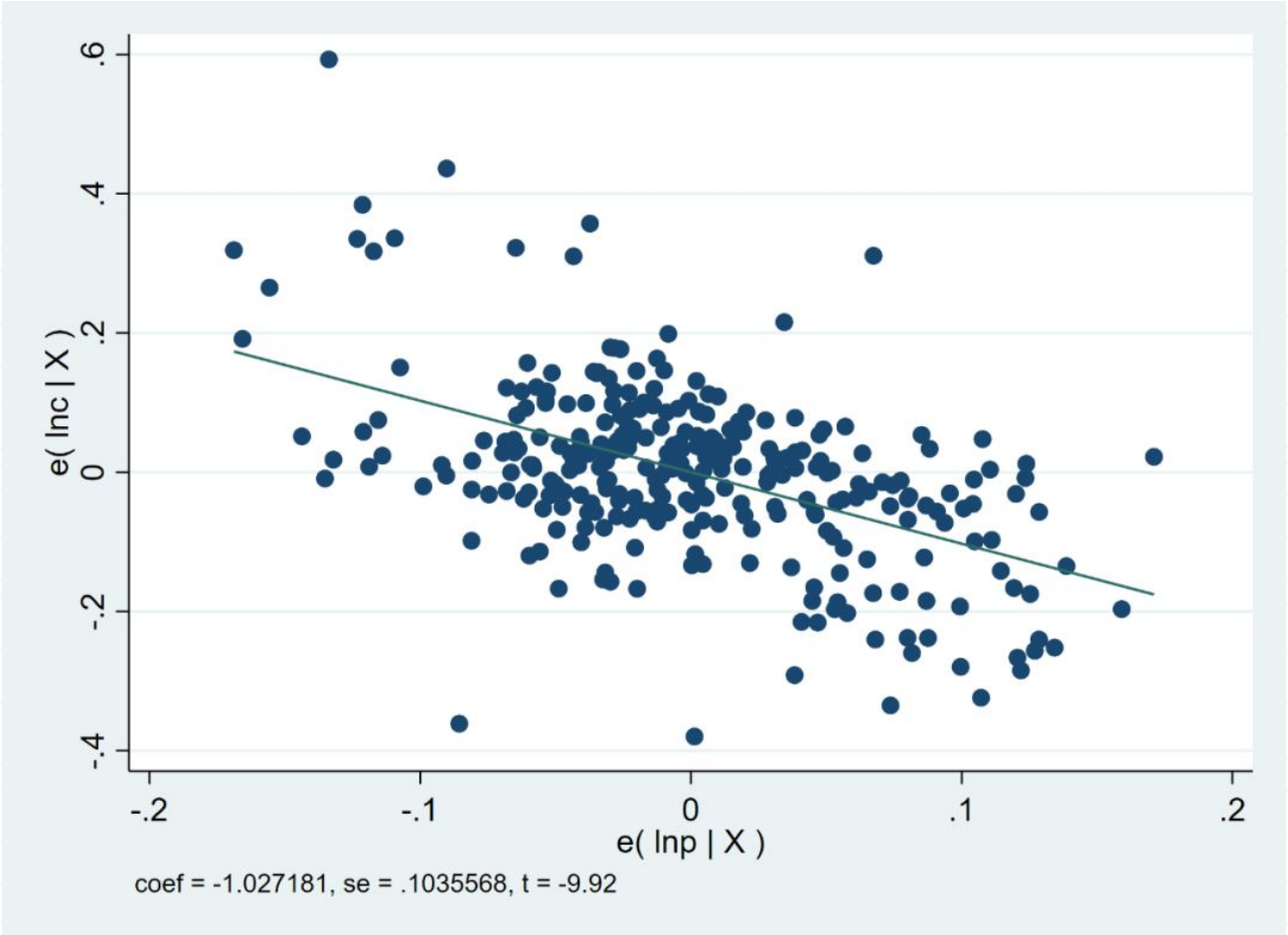
输入命令【est store ols】保存结果。

```
. reg lnc lnp lnpmin lny state2-state10 t
```

Source	SS	df	MS	Number of obs	=	300
				F(13, 286)	=	56.65
Model	9.24427482	13	.711098063	Prob > F	=	0.0000
Residual	3.58977229	286	.012551651	R-squared	=	0.7203
				Adj R-squared	=	0.7076
Total	12.8340471	299	.042923234	Root MSE	=	.11203

lnc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnp	-1.027181	.1035568	-9.92	0.000	-1.231011	-.8233509
lnpmin	.5100582	.101909	5.01	0.000	.3094714	.710645
lny	.4975365	.152624	3.26	0.001	.1971278	.7979453
state2	-.0773908	.0384839	-2.01	0.045	-.1531385	-.0016432
state3	.088557	.029409	3.01	0.003	.0306715	.1464424
state4	-.1809375	.0712344	-2.54	0.012	-.3211478	-.0407272
state5	-.1066138	.0888135	-1.20	0.231	-.2814249	.0681973
state6	.2177434	.0476398	4.57	0.000	.1239743	.3115124
state7	.115543	.0750084	1.54	0.125	-.0320954	.2631815
state8	.1068277	.0492755	2.17	0.031	.0098391	.2038163
state9	.0433207	.0328848	1.32	0.189	-.0214061	.1080476
state10	-.133583	.0328065	-4.07	0.000	-.1981558	-.0690101
t	-.0429824	.0120418	-3.57	0.000	-.0666841	-.0192807
_cons	3.488747	1.377469	2.53	0.012	.777485	6.20001

然后，我们输入命令【avplot lnp】，查看核心解释变量lnp与被解释变量lnc的偏相关图。



可以发现，两者之间明显呈现负相关关系。

接下来，我们检验误差项是否存在自相关、异方差和截面相关的问题。

1、自相关的检验

使用命令【xtserial lnc lnp lnpmin lny state2-state10 t】检验误差项的自相关问题。

```
. xtserial lnc lnp lnpmin lny state2-state10 t


Wooldridge test for autocorrelation in panel data
H0: no first-order autocorrelation
       F(  1,         9) =      89.304
            Prob > F =       0.0000
```

由检验结果可知，P值为0，所以拒绝一阶自相关不存在的原假设，表明存在自相关问题。

2、异方差的检验

使用命令【xttest3】对误差项的异方差问题进行检验。

#【xttest3】只能在【xtreg,fe】和【xtgls】命令之后使用；

#第一次使用【xttest3】的同学，需要使用命令【ssc install xttest3】进行安装

在Stata中输入命令【quietly xtreg lnc lnp lnpmin lny t,fe】,然后输入命令【xttest3】。

（也可以输入命令【quietly xtgls lnc lnp lnpmin lny state2-state10 t】和【xttest3】）

```
. quietly xtreg lnc lnp lnpmin lny t,fe

. xttest3

Modified Wald test for groupwise heteroskedasticity
in fixed effect regression model

H0: sigma(i)^2 = sigma^2 for all i

chi2 (10)  =      378.90
Prob>chi2 =      0.0000
```

由检验结果可知，P值为0，所以拒绝原假设，认为误差项存在异方差问题。

3、截面相关的检验

使用命令【xttest2】对误差项的截面相关问题进行检验。

#【xttest2】只能在【xtreg,fe】、【xtgls】或【ivreg2】之后使用，只适用于长面板数据；

#第一次使用【xttest2】的同学，需要使用命令【ssc install xttest2】进行安装

在Stata中输入命令【quietly xtreg lnc lnp lnpmin lny t,fe】,然后输入命令【xttest2】。

（也可以输入命令【quietly xtgls lnc lnp lnpmin lny state2-state10 t】和【xttest2】）

当然，因为我们上一步进行了误差项异方差问题的检验，所以这一步我们可以直接输入命令【xttest2】。

```
. xttest2

Correlation matrix of residuals:

           __e1     __e2     __e3     __e4     __e5     __e6     __e7     __e8     __e9     __e10
 __e1   1.0000
 __e2  -0.0937   1.0000
 __e3   0.9592  -0.0621   1.0000
 __e4  -0.4242   0.3875  -0.4670   1.0000
 __e5  -0.5426   0.3441  -0.5872   0.5519   1.0000
 __e6   0.0245   0.5696  -0.0405   0.5177   0.5805   1.0000
 __e7  -0.7434   0.4153  -0.7509   0.5701   0.8446   0.4893   1.0000
 __e8   0.5650   0.5380   0.5281   0.1007  -0.2150   0.4899  -0.3263   1.0000
 __e9   0.8337   0.2859   0.8507  -0.2972  -0.3914   0.1548  -0.5800   0.7129   1.0000
__e10   0.7510   0.3314   0.7628   0.0002  -0.1575   0.2508  -0.4293   0.6318   0.8345   1.0000

Breusch-Pagan LM test of independence: chi2(45) =   376.963, Pr = 0.0000
Based on 30 complete observations over panel units
```

可以看到，检验结果的P值为0，所以拒绝原假设，认为误差项存在截面相关的问题。

综上，通过检验，我们发现模型误差项存在自相关、异方差和截面相关的问题。

第四步 报告计量结果

通过第三步对模型误差项的检验，我们知道模型的误差项存在自相关、异方差和截面相关的问题，所以，我们需要对误差项的自相关、异方差和截面相关问题进行处理并报告计量结果。

对【xtpcse】、【xtgls】和【xtscc】三个命令的结果分别进行报告。

依次输入命令：

【xtpcse lnc lnp lnpmin lny state2-state10 t,corr(psar1)】

【est store xtpcse】

【xtgls lnc lnp lnpmin lny state2-state10 t,corr(psar1) panels(correlated)】

【est store xtgls】

【xtscc lnc lnp lnpmin lny state2-state10 t】

【est store xtscc】

最后通过【esttab】命令将所有的存储结果放在一起进行比较。

输入命令【esttab ols xtpcse xtgls xtscc,b(%9.2f)p mtitle(ols xtpcse xtgls xtscc)obslast star(* 0.1 ** 0.05 *** 0.01)compress nogap k(lnp lnpmin lny t)】

	(1) ols	(2) xtpcse	(3) xtgls	(4) xtscc
lnp	-1.03***	-0.30***	-0.35***	-1.03***
	(0.000)	(0.000)	(0.000)	(0.000)
lnpmin	0.51***	0.05	0.02	0.51**
	(0.000)	(0.451)	(0.541)	(0.011)
lny	0.50***	0.53***	0.55***	0.50**
	(0.001)	(0.000)	(0.000)	(0.015)
t	-0.04***	-0.05***	-0.05***	-0.04***
	(0.000)	(0.000)	(0.000)	(0.005)
N	300	300	300	300

p-values in parentheses
* p<0.1, ** p<0.05, *** p<0.01

输出的表格中，(1)的结果是不对误差项做任何处理的结果，(2)、(3)、(4)是分别使用三种命令并对误差项的三大问题进行处理的结果。