# Estimating partially linear **functional-coefficient** panel data models with Stata

杜克锐

厦门大学 管理学院

# 关于交互项的那些事（一）

原创：陈强 计量经济学及Stata应用 3月2日

计量经济学的初学者经常困惑于交互项的含义，以及在实证研究中如何耳
实上，如果善于使用交互项，或许可以为你的论文增色不少。本文就来耶
的那些事。

**什么是交互项？**

考虑简单的二元线性回归模型：

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

# 关于交互项的那些事（二）：画交互效应图原来如此简单

原创：陈强 计量经济学及Stata应用 3月24日

在上期推文中，我们从边际效应（求偏导数）的角度考察了交互效应的经济含义。这隐含地假设了，在交互项的两个变量中，至少有一个为连续变量，故可以求偏导数。显然，如果两个变量都是离散变量（比如虚拟变量），则此法不通，因为无法求导数。

但离散变量的交互效应也有其方便之处，可以将相应的回归方程用分段函数来表示。比如，考虑如下包含交互项的模型：

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i$$

# 关于交互项的那些事（三）：遗漏变量偏差 vs 多重共线性

原创：陈强 计量经济学及Stata应用 6月8日

纯粹的线性模型只是对现实世界的近似，类似于一阶泰勒展开。将线性模型推广至非线性模型的方法很多。最简单的做法就是加入平方项与交互项，这相当于二阶泰勒展开。

从数学上看，二阶泰勒展开对于（真实的）非线性函数的近似效果肯定比一次函数更好。但为什么在经济学的应用中却很少看到平方项与交互项呢？即使有交互项，也通常只是两个变量的之间的交互项，为何不考虑所有变量之间的各种可能的交互项呢？这涉及到需要在“遗漏变量偏差”（omitted variable bias）与“多重共线性”（multicollinearity）之间作权衡。

# 关于交互项的那些事（四）：非线性模型中的交互效应

原创：陈强 计量经济学及Stata应用 7月5日

如果交互效应出现在非线性模型中，比如二值选择模型（Probit或Logit），应该
释呢？与线性模型中的交互效应有何不同？

事实上，二者很不相同！艾春荣老师最早注意到了这个“奇怪”的现象（Ai and N
2003）。

## Stata: 交叉项\交乘项该这么分析！

原创： Stata连享会 Stata连享会 5月27日

> Source： Jens Hainmueller, Jonathan Mummolo and Yiqing Xu, 2018, How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice
> 原文作者： Jens Hainmueller, Jonathan Mummolo and Yiqing Xu
> 原文链接： http://yiqingxu.org/papers/english/2018_HMX_interaction/main.pdf

> 推文作者： 刘聪聪 (中山大学)，陈点点 (中国人民大学)
> Stata 连享会： 知乎 | 简书 | 码云

> 特别说明

## Stata - 图示调节效应

原创： Stata连享会 Stata连享会 3月22日

> 作者： 邓浩然 (上海财经大学)
> Stata 连享会： 知乎 | 简书 | 码云

# Interaction Effect

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i$$

$$\frac{\partial y_i}{\partial x_{1i}} = \beta_1 + \beta_3 x_{i2}$$

# Interaction Effect

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i$$

$$\frac{\partial y_i}{\partial x_{1i}} = \beta_1 + \beta_3 x_{i2}$$

$$\frac{\partial y_i}{\partial x_{1i}} = g(x_{i2})$$

# Partially Linear Functional-coefficient Panel Data Model

$$Y_{it} = Z_{it}^{'} g\left(U_{it}\right) + X_{it}^{'} \beta + a_i + \varepsilon_{it}, i = 1, \ldots, N, t = 1, \ldots, T$$

# Partially Linear Functional-coefficient Panel Data Model

$$Y_{it} = Z_{it}^{'} g\left(U_{it}\right) + X_{it}^{'} \beta + a_i + \varepsilon_{it}, i = 1, \ldots, N, t = 1, \ldots, T$$

$$Z_{it} = \left(Z_{1,it}, \ldots, Z_{l,it}\right)^{'}; U_{it} = \left(U_{1,it}, \ldots, U_{l,it}\right)^{'}; g\left(U_{it}\right) = \left(g_1\left(U_{1,it}\right), \ldots, g_l\left(U_{l,it}\right)\right)^{'}$$

$$Z_{it}^{'} g\left(U_{it}\right) = \sum_{j=1}^{l} Z_{j,it}^{'} g_j\left(U_{j,it}\right)$$

# Partially Linear Functional-coefficient Panel Data Model

$$Y_{it} = Z_{it}^{'} g\left(U_{it}\right) + X_{it}^{'}\beta + a_i + \varepsilon_{it}, i = 1,\ldots,N, t = 1,\ldots,T$$

$$U_{it} = \left(U_{1,it},\ldots,U_{l,it}\right)^{'}; g\left(U_{it}\right) = \left(g_1\left(U_{1,it}\right),\ldots,g_l\left(U_{l,it}\right)\right)^{'}$$

$Z_{it}$ is a scalar random variable; $Z_{it}^{'} g\left(U_{it}\right) = Z_{it}^{'}\left(\sum_{j=1}^{l} g_j\left(U_{j,it}\right)\right)$

$$Z_{it} = \left(Z_{1,it},\ldots,Z_{l,it}\right)^{'};$$ $U_{it}$ is a scalar random variable

$$Z_{it}^{'} g\left(U_{it}\right) = \sum_{j=1}^{l} Z_{j,it}^{'} g_j\left(U_{it}\right)$$

# Partially Linear Functional-coefficient Panel Data Model

$$Y_{it} = Z_{it}^{'} g\left(U_{it}\right) + X_{it}^{'}\beta + a_i + \varepsilon_{it}, i = 1,\ldots,N, t = 1,\ldots,T$$

$$Z_{it}^{'} \triangleq 1$$

$$Y_{it} = g\left(U_{it}\right) + X_{it}^{'}\beta + a_i + \varepsilon_{it}, i = 1,\ldots,N, t = 1,\ldots,T$$

Panel data version of ***npregress series***

# Estimation Procedure

- An et al. (2016) and Zhang and Zhou (2018) propose using the series method to estimate the model.
- Step 1. Using a linear combination of sieve basis functions to approximate the unknown functional coefficients

$$h(U_{it}) = \left(h_1(U_{it}), \ldots, h_L(U_{it})\right)'$$

$$g(U_{it}) \approx h(U_{it})'\gamma$$

$$Y_{it} = H_{it}'\gamma + X_{it}'\beta + a_i + v_{it}$$

$$H_{it} \equiv \left(Z_{it}h(U_{it})', \ldots, Z_{it}h(U_{it})'\right)'$$

# Estimation Procedure

- An et al. (2016) and Zhang and Zhou (2018) propose using the series method to estimate the model.
- Step 2. Taking the first time difference of model to eliminate the fixed effects.

$$\Delta Y_{it} = \Delta H_{it}^{'} \gamma + \Delta X_{it}^{'} \beta + \Delta v_{it}$$

# Estimation Procedure

- An et al. (2016) and Zhang and Zhou (2018) propose using the series method to estimate the model.
- Step 3. Estimating the model via least square(LS) method
- Step 4. Predict the unknown functional coefficients

$$\hat{g}_p\left(U_{it}\right) = h\left(U_{it}\right)' \hat{\gamma}_p$$

Remarks: Under certain regular assumptions, Zhang and Zhou (2018) establish the consistency and asymptotic normality of the above estimator when sample size N and T go to infinity jointly or only N tends to infinity.

# Stata commands

① **xtplfc** estimates An et al.'s (2016) partially linear functional-coefficient panel data models with exogenous variables.

② **ivxtplfc** estimates Zhang and Zhou's (2018) partially linear functional-coefficient panel data models with endogenous variables using the sieve 2SLS method.

③ **xtdplfc** estimates Zhang and Zhou's (2018) partially linear functional-coefficient panel dynamic data models using the sieve 2SLS method.

# Stata commands

```
xtplfc varlist, zvars(varlist) uvars(varname) generate(string) [
    power(numlist) nknots(numlist) quantile maxnknots(numlist)
    minnknots(numlist) grid(string) pctile(#) brep(#) wild
    predict(prspec) nodots level(#) fast cvtwo ]
```

# Stata commands

```
ivxtplfc varlist, zvars(varlist) uvars(varlist) generate(string) [
    endox(varlist) endozflag(numlist) ivx(varlist) ivz(varlist,
    uflag(numlist) [ivtype(#)]) power(numlist) nknots(numlist)
    quantile(numlist) maxnknots(numlist) maxnknots(numlist)
    grid(string) pctile(#) brep(#) wild predict(prspec) nodots

    level(#) fast cvtwo ]
```

# Stata commands

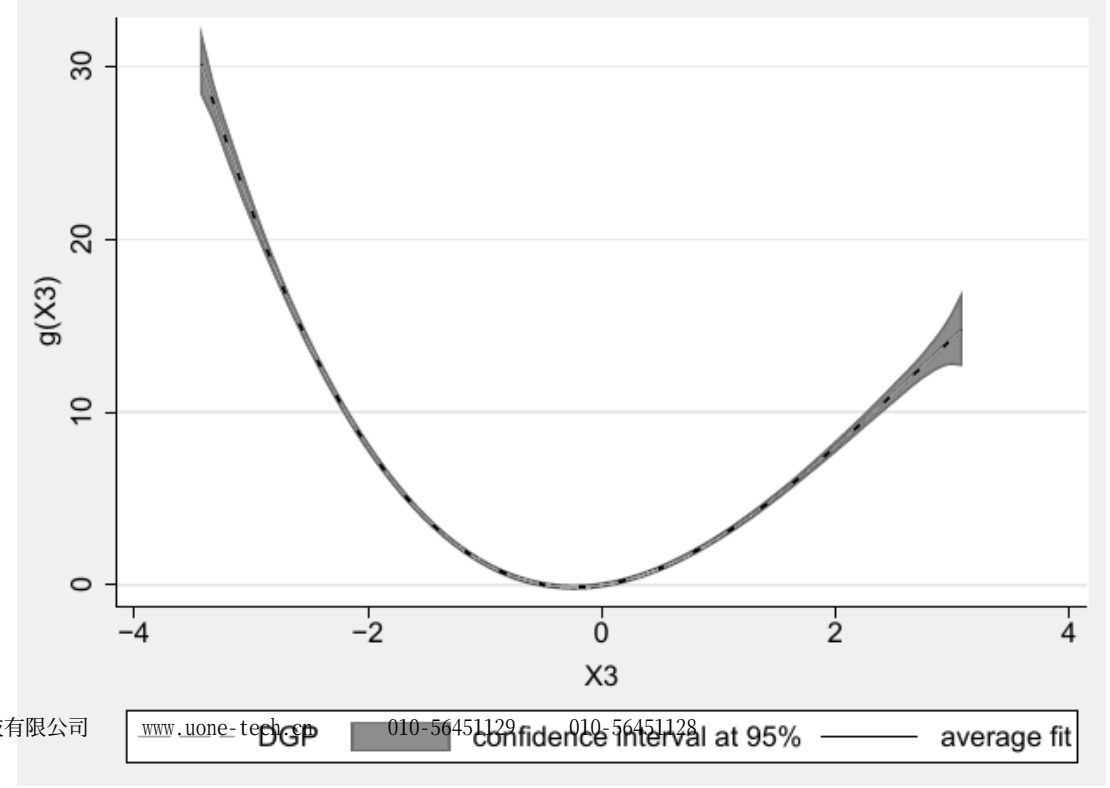
```
xtdplfc varlist, uvars(varlist) generate(string) [ zvars(varlist)
    lag(#) lagyinz(numlist) endox(varlist) endozflag(numlist)
    ivx(varlist) ivz(varlist, uflag(numlist) [ivtype(#)]) onlyivxz
    ivtype(numlist) power(numlist) nknots(numlist)
    quantile(numlist) maxnknots(numlist) maxnknots(numlist)
    grid(string) pctile(#) brep(#) wild predict(prspec) nodots
    level(#) fast cvtwo ]
```

# Simulations

DGP 1: $$Y_{it} = X_{1,it} - X_{2,it} + g\left(X_{3,it}\right)Z_{it} + a_i + \varepsilon_{it}$$

$$g\left(X_{3,it}\right) = X_{3,it} + 2\left(X_{3,it}\right)^2 - 0.25\left(X_{3,it}\right)^3$$

N=50, T=40



# Simulations

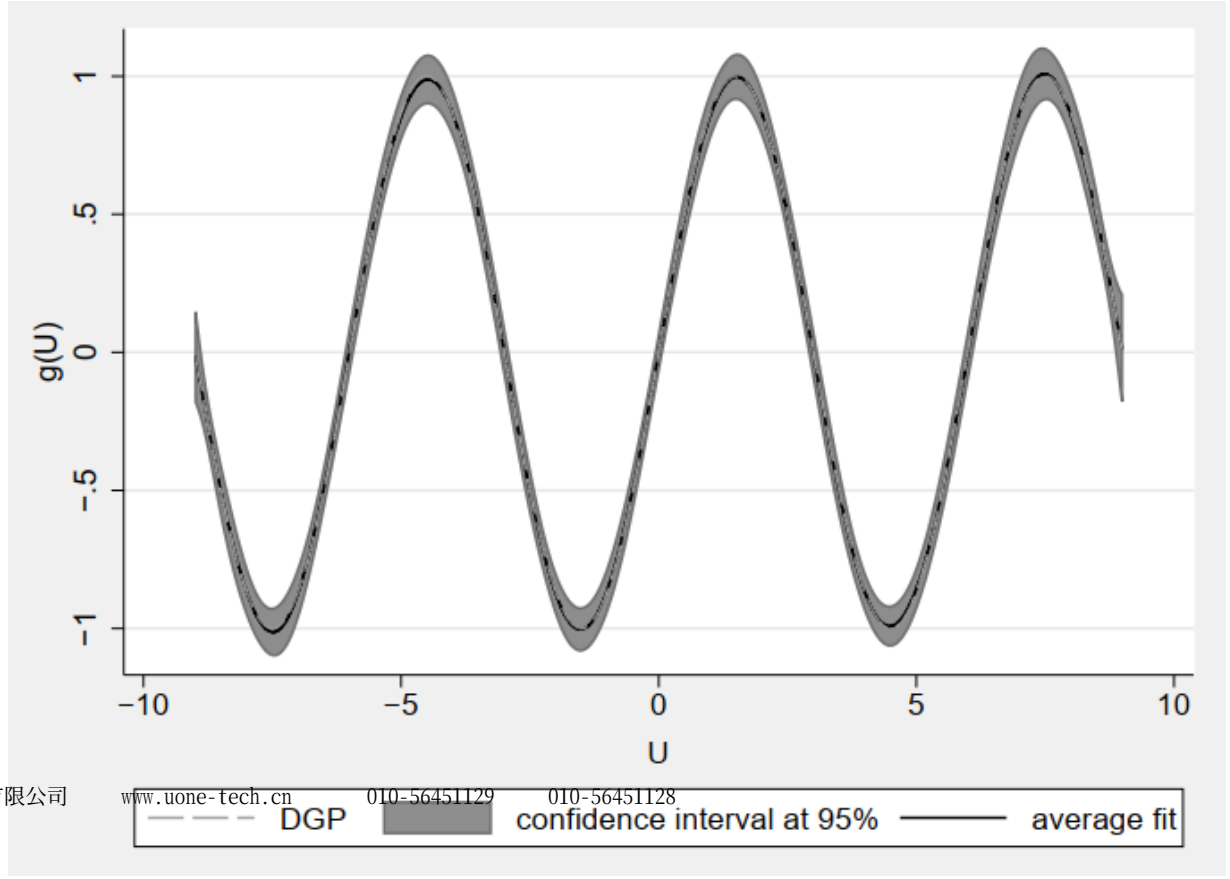
DGP 2:

$$Y_{it} = g(U_{it})Y_{it-1} + 0.3Y_{it-2} + 0.3X_{it} + a_i + \varepsilon_{it}$$

$$g(U_{it}) = \sin\left(\frac{\pi}{3}U_{it}\right)$$

$$X_{it} = W_{it} + 0.5a_i$$

N=50, T=40



# Applications

- https://www.sciencedirect.com/science/article/pii/S1043951X19300574

# Concluding remarks

## Heterogenous Analysis

【铭声音】实证研究"13"点-9-为什么要去做异质性分析?

铭心而论 7月3日

无题

陆铭【摄】

实证研究"13"点-9-为什么要

来自铭心而论

15:18

## Exploring the determinants of the impact

- Regress Y on X, and the coefficient of X, β captures the impact of X on Y
- PLFC models investigate the impact of some variables of interest on β (i.e., the impact of X on Y)

# Concluding remarks

Untitled.do*

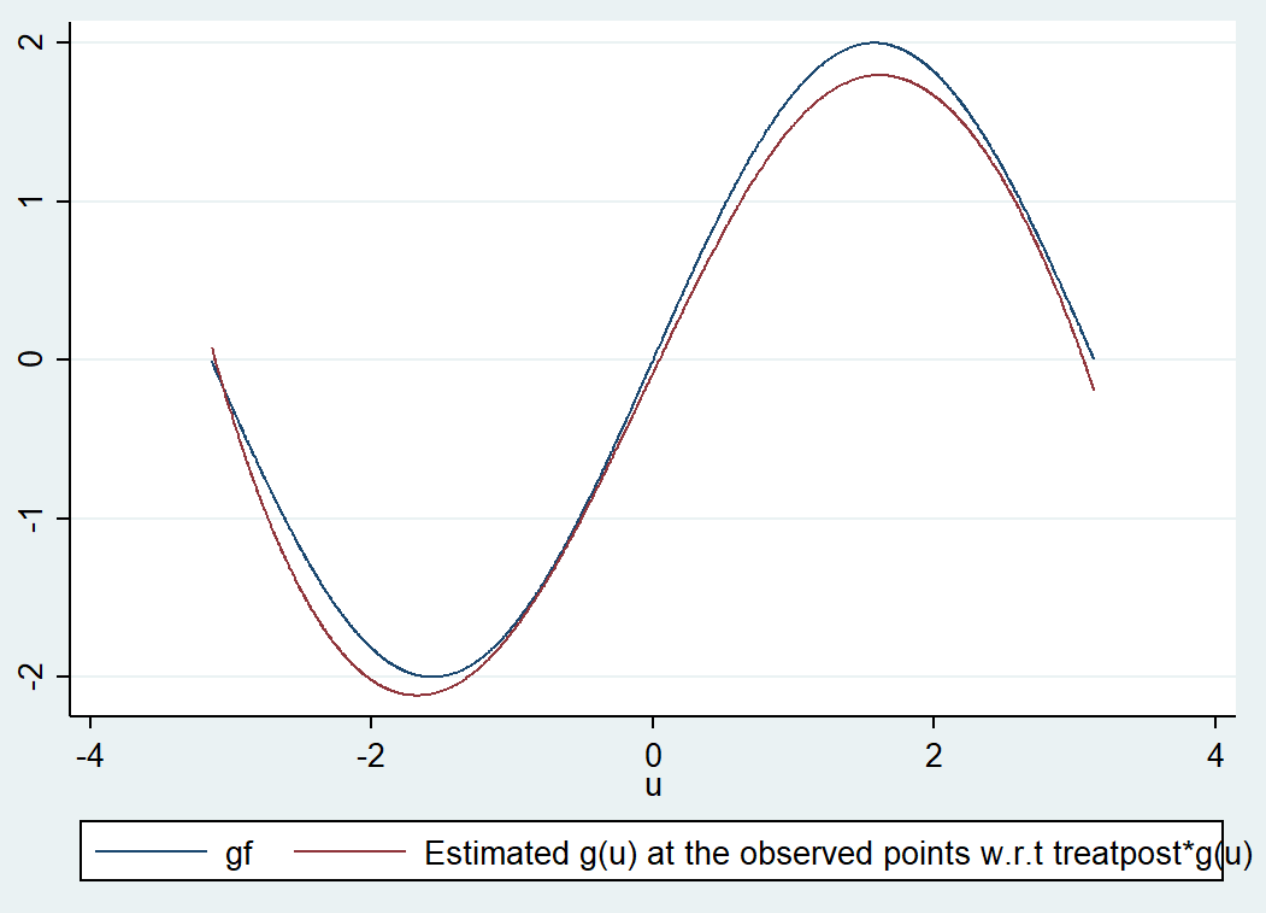
```
clear all
set obs 500
gen a=rnormal()
gen treat=1 if _n<=400
replace treat=0 if missing(treat)
gen id=_n
expand 10
bys id: gen year=_n
gen post=0
replace post=1 if year>=5
gen x=2*runiform()
gen u=-_pi+2*_pi*runiform()
gen gf=2*sin(u)
gen y=a+2*x+2*post+treat*post*gf+sqrt(2)*rnormal()
xtset id year
gen treatpost=treat*post

xtplfc y x  post, z(treatpost) u(u) maxnk(10) gen(f) fast
twoway (line gf u if treatpost==1,sort) (line f_1 u if treatpost==1,sort)


xtreg y x post treatpost, fe r

gen hte=u*treatpost
xtreg y x post treatpost hte, fe r
```

# Concluding remarks



# Concluding remarks

```
.
. xtreg y x post treatpost, fe r

Fixed-effects (within) regression               Number of obs      =         5,000
Group variable: id                              Number of groups   =           500

R-sq:                                           Obs per group:
     within  = 0.4561                                         min =            10
     between = 0.1206                                         avg =          10.0
     overall = 0.3778                                         max =            10

                                                F(3,499)           =       1383.47
corr(u_i, Xb)  = 0.0112                         Prob > F           =        0.0000

                                        (Std. Err. adjusted for 500 clusters in id)
------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    2.065947   .0457277    45.18   0.000     1.976104    2.155789
        post |    1.962503   .0978968    20.05   0.000     1.770162    2.154844
   treatpost |    .1102983   .1103663     1.00   0.318    -.1065416    .3271382
       _cons |    .0258731   .0527916     0.49   0.624     -.077848    .1295942
-------------+----------------------------------------------------------------
     sigma_u |  1.1495198
     sigma_e |   1.730536
         rho |   .30615146   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

# Concluding remarks

ow Help

```
. gen hte=u*treatpost

. xtreg y x post treatpost hte, fe r

Fixed-effects (within) regression               Number of obs      =         5,000
Group variable: id                              Number of groups   =           500

R-sq:                                           Obs per group:
     within  = 0.5628                                         min =            10
     between = 0.1690                                         avg =          10.0
     overall = 0.4710                                         max =            10

                                                F(4,499)           =       1272.37
corr(u_i, Xb)  = 0.0132                         Prob > F           =        0.0000

                                  (Std. Err. adjusted for 500 clusters in id)
------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    2.040685   .0421022    48.47   0.000     1.957965    2.123404
        post |    1.963303   .0977745    20.08   0.000     1.771203    2.155403
   treatpost |    .0625498   .1081001     0.58   0.563    -.1498376    .2749372
         hte |    .6137152   .0218934    28.03   0.000     .5707007    .6567297
       _cons |    .0513001   .0500182     1.03   0.306    -.0469722    .1495724
-------------+----------------------------------------------------------------
     sigma_u |  1.1176926
     sigma_e |   1.551592
         rho |   .34163192   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

# Concluding remarks

## Heterogenous Analysis

【铭声音】实证研究"13"点-9-为什么要去做异质性分析?

铭心而论 7月3日

无题

陆铭【摄】

实证研究"13"点-9-为什么要

来自铭心而论

15:18

## Exploring the determinants of the impact

- Regress Y on X, and the coefficient of X, β captures the impact of X on Y
- PLFC models investigate the impact of some variables of interest on β (i.e., the impact of X on Y)

# 谢谢

# Thank You