《数据分析与建模》实验课程教学案例

撰写人：吴伟

2017 年 9 月 25 日

目录

数据分析与建模案例集

案例一：数据管理

注：吴伟改编，对外经济贸易大学，本案例只应用于教学目标

摘要：本案例主要对数据管理进行练习

关键词：数据清洁、Stata

背景描述：应用 Stata 进行数据清理工作，这涉及到数据管理工作，是进行统计与计量分析的重要前提。

1.教学目的与要求：适用课程:《数据分析与建模》、《公共管理方法与技术》、《社会经济统计学》，本科经济学、管理学

2.启发思考题：使用数据自己进行操作分析

3.应用 stata13.1 进行数据管理

4.代码

5.建议学时，3-4 学时

```
//    experiment: conduct data using stata,
//    Wei We, Department of Public Economics, SPA, UIBE

version 11.2
clear all
macro drop _all
set linesize 80
set scheme s2manual
program drop _all

use binlfp4, clear

tabulate hc wc, row nolabel
tabulate hc wc, /// the next line is treated as a continuation of this one
      row nolabel
```

```
//    Syntax of Stata commands

tabulate hc wc if age>40, row
summarize inc k5 wc
summarize
summarize inc if age>=25 & age<=65
summarize inc if (age<25 | age>65) & age<. & hc==1

* getting information about variables

codebook lfp k5 k618 agecat wc hc lwg inc, compact
summarize inc, detail
tabulate hc
tabulate hc, nolabel
tabulate hc wc
tab1 hc wc

dotplot inc
graph export rm3ch2-dotplot.emf, replace
describe lfp k5 k618 agecat wc hc lwg inc
codebook inc

generate age2 = age
summarize age2 age

gen age3 = age if age>40
gen agesq = age^2
gen lnage = ln(age)

* replace command
```

```
gen age4 = age
replace age4 = 40 if age<40
summarize age4 age

* recode command

use recodedata4, clear // note changed dataset name
recode origvar (1=2) (3=4), generate(myvar1)
recode origvar (2=1) (*=0), generate(myvar2)
recode origvar (2=1) (nonmissing=0), generate(myvar3)
recode origvar (1/4=2), generate(myvar4)
recode origvar (1 3 4 5=7), generate(myvar5)
recode origvar (min/5=min), generate(myvar6)
recode origvar (missing=9), generate(myvar7)

//    Labeling variables and values

use gsskidvalue4, clear
gen agesq = age*age
label variable agesq "age-squared of respondent"
codebook age agesq, compact
label variable agesq
codebook agesq, compact
label define Lyesno 1 yes 0 no
label define Lposneg4 1 veryN 2 negative 3 positive 4 veryP
label define Lagree4 1 StrongA 2 Agree 3 Disagree 4 StrongD
label define Lagree5 1 StrongA 2 Agree 3 Neutral 4 Disagree 5 StrongD

label values female Lyesno
```

```
label values black Lyesno
label values anykids Lyesno
describe female black anykids

tabulate anykids
label define degree 0 "no_hs" 1 "hs" 2 "jun_col" 3 "bachelor" 4 "graduate"
label values degree degree
tabulate degree

notes: General Social Survey extract for Stata book | J Freese | 2014-01-23
notes income: self-reported family income, measured in dollars
notes income: refusals coded as missing
notes

//    Global and local macros

use binlfp4, clear
global myoptions ", cell miss nolabel chi2 nokey"
tabulate lfp wc $myoptions
tabulate lfp wc, cell miss nolabel chi2 nokey

local myoptions ", cell miss nolabel chi2 nokey"
tab lfp wc `myoptions'
local demogvars "age white female"
local edvars "highsch college graddeg"
di "regress y `demogvars' `edvars' x1 x2 x3"
di "regress y age white female highsch college graddeg x1 x2 x3"

local wclabel : variable label wc
display "`wclabel'"
```

```
//    Loops using foreach and forvalues

/* examples of loop grammar

foreach cutpt in 2 3 4 {
        generate y_lt`cutpt' = y<`cutpt' if y<.
}

generate y_lt2 = y<2 if y<.
local rhs "yr89 male white age ed prst"

foreach lhs in y_lt2 y_lt3 y_lt4 {
        logit `lhs' `rhs'
}

foreach lhs in y_lt2 y_lt3 y_lt4 {
        tabulate `lhs'
        logit `lhs' `rhs'
        probit `lhs' `rhs'
}

*/

//    Graphics

use lfpgraph4, clear
codebook income agecat1pr1 agecat2pr1 agecat3pr1, compact
list income agecat1pr1 agecat2pr1 agecat3pr1
```

```
graph twoway scatter agecat1pr1 agecat2pr1 agecat3pr1 income, ///
      ytitle(Probability)
graph export rm3ch2-scatter.emf, replace

graph twoway (connected agecat1pr1 income) ///
      (scatter agecat2pr1 agecat3pr1 income), ytitle(Probability)
graph export rm3ch2-connected.emf, replace

graph twoway connected agecat1pr1 agecat2pr1 agecat3pr1 income, ///
      title("Predicted Probability of Female LFP") ///
      subtitle("(as predicted by logit model)") ///
      ytitle("Probability") xtitle("Family income, excluding wife's") ///
      caption("Data from 1976 PSID compiled by T Mroz")
graph export rm3ch2-titles.emf, replace

graph twoway connected agecat1pr1 agecat2pr1 agecat3pr1 income, ///
      title("Predicted Probability of Female LFP") ///
      subtitle("(as predicted by logit model)") ///
      ytitle("Probability") xtitle("Family income, excluding wife's") ///
      caption("Data from 1976 PSID compiled by T Mroz") ///
      xlabel(10 20 30 40 50 60 70 80 90)
graph export rm3ch2-axes.emf, replace

graph twoway connected agecat1pr1 agecat2pr1 agecat3pr1 income, ///
      title("Predicted Probability of Female LFP") ///
      subtitle("(as predicted by logit model)") ///
      ytitle("Probability") xtitle("Family income, excluding wife's") ///
      caption("Data from 1976 PSID compiled by T Mroz") ///
      xlabel(10(10)90) name(graph1, replace)
graph export rm3ch2-names.emf, replace
```

```
* combining graphs

use gssclass4, clear
ologit class i.female i.white i.year i.educ c.age##c.age income, nolog

mgen, at(income=(0(25)250)) stub(CL_) atmeans
label var CL_pr1 "Lower"
label var CL_pr2 "Working"
label var CL_pr3 "Middle"
label var CL_pr4 "Upper"
label var CL_Cpr1 "Lower"
label var CL_Cpr2 "Lower/Working"
label var CL_Cpr3 "Lower/Working/Middle"

graph twoway connected CL_pr1 CL_pr2 CL_pr3 CL_pr4 CL_income, ///
      title("Panel A: Predicted Probabilities") ///
      xtitle("Household income (2012 dollars)") ///
      xlabel(0(50)250) ylabel(0(.25)1, grid gmin gmax) ///
      xline(68.1, lpattern(dash)) ytitle("") name(panelA, replace)

graph twoway connected CL_Cpr1 CL_Cpr2 CL_Cpr3 CL_income, ///
      title("Panel B: Cumulative Probabilities") ///
      xtitle("Household income (2012 dollars)") ///
      xlabel(0(50)250) ylabel(0(.25)1, grid gmin gmax) ///
      xline(68.2, lpattern(dash)) ytitle("") name(panelB, replace)

graph combine panelA panelB, xsize(8) ysize(4) ///
      caption("Example of combining horizontally.")
graph export rm3ch2-combine-horiz.emf, replace
```

```
graph combine panelA panelB, col(1) xsize(4) ysize(6) ///
      caption("Example of combining vertically.")
graph export rm3ch2-combine-vert.emf, replace

log close
exit
```

案例二、美国的电力消费

注：吴伟改编，对外经济贸易大学，本案例只应用于教学目标

摘要：本案例主要对数据进行概括性统计及做表格

关键词：描述性统计分析、统计表格、Stata

背景描述：应用 Stata 进行数据描述性统计，是进行计量分析的第一步。

1.教学目的与要求：适用课程:《数据分析与建模》、《公共管理方法与技术》、《社会经济统计学》，本科经济学、管理学

2.启发思考题：使用数据自己进行操作分析

3.应用 stata13.1 进行描述性统计分析

4.代码

5.建议学时，2 学时

```
//exercise on the descriptive statistics and statistical table

Use electricity.dta, clear
describe
summarize elcap
summarize elcap, detail
tabstat elcap, stats(mean min max)
            tabstat elcap, stats(mean min max) by(region4) // constructs a
      table containing summary statistics /// for each value of varname
```

```
tabstat elcap, stats(mean min max) by(region4)
codebook, compact
```

*produce a 99% confidence interval for inference of population (kWh per person) in the U.S.

```
ci elcap, level(99)
```

案例三、美国的能源消费

注：吴伟改编，对外经济贸易大学，本案例只应用于教学目标

摘要：本案例主要用数据进行方差分析，练习初步的政策评价方法

关键词：方差分析、协方差分析、Stata

背景描述：应用 Stata 进行数据方差分析，是应用统计方法进行政策评价的基础方法。

1.教学目的与要求：适用课程:《数据分析与建模》、《公共管理方法与技术》、《社会经济统计学》,使学生理解方差分析的思想和原理。公共管理专业本科生，使学生理解如何比较两个以上的总体均值的原理。面向本科经济学、管理学。

2.启发思考题：使用数据自己进行操作分析

3.应用 stata13.1 进行方差分析

4.代码

5.建议学时，2 学时

见数据 energy consumption.sav

美国能源消费局收集了居民的能源消耗和支出。假定我们想确定在美国四个区域中平均家庭年度消费是否具有差异性。假定$\mu_1, \mu_2, \mu_3, \mu_4$分别表示美国东北、中西、南部和西部家庭上一年度的能源消费。待检验假设为：

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$(平均能源消费都相等)

H_1: 不是所有均值都相等(如至少有两个均值不相等)

试根据数据检验上述均值是否相等。

Descriptive Statistics

Dependent Variable: 年度能源消费

region	Mean	Std. Deviation	N
northeast	11.00	1.871	5
midwest	12.50	2.588	6
south	7.50	3.000	4
west	7.20	1.924	5
Total	9.80	3.205	20

Tests of Between-Subjects Effects

Dependent Variable: 年度能源消费

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	105.900[a]	3	35.300	6.325	.005
Intercept	1786.824	1	1786.824	320.148	.000
region	105.900	3	35.300	6.325	.005
Error	89.300	16	5.581		
Total	2116.000	20			
Corrected Total	195.200	19			

a. R Squared = .543 (Adjusted R Squared = .457)

表明能源消费具有差异。

案例四、政府工作项目有效性

注：吴伟改编，对外经济贸易大学，本案例只应用于教学目标

摘要：本案例主要用数据进行协方差分析，练习初步的政策评价方法

关键词：协方差分析、Stata

背景描述：应用 Stata 进行数据协方差分析，是应用统计方法进行政策评价的基础方法。

1.教学目的与要求：适用课程：《数据分析与建模》、《公共管理方法与技术》、《社会经济统计学》，使学生理解协方差分析的思想和原理。公共管理专业本科生，面向本科经济学、管理学。

2.启发思考题：使用数据自己进行操作分析

3.应用 stata13.1 进行方差分析

4.代码

5.建议学时，3 学时

数据：workprog.sav

支持政府工作计划的人想看看是否该计划帮助人们找到了更好的工作，在进入培训计划时需要控制薪水水平。计划的潜在参与者样本确定后，随机选取潜在人选来参加政府工作计划，其他的不参加计划。

Tests of Between-Subjects Effects

Dependent Variable: Income after the program

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	12295.033[a]	3	4098.344	429.755	.000
Intercept	131.271	1	131.271	13.765	.000
prog	106.795	1	106.795	11.199	.001
incbef	7152.586	1	7152.586	750.025	.000
prog * incbef	4.292	1	4.292	.450	.502
Error	9498.318	996	9.536		
Total	297121.000	1000			
Corrected Total	21793.351	999			

a. R Squared = .564 (Adjusted R Squared = .563)

交互项不显著，参加培训与未参加培训人的之前收入无差异。

Tests of Between-Subjects Effects

Dependent Variable: Income after the program

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	12290.741[a]	2	6145.370	644.763	.000
Intercept	131.400	1	131.400	13.786	.000
prog	4735.662	1	4735.662	496.859	.000
incbef	7153.844	1	7153.844	750.571	.000
Error	9502.610	997	9.531		
Total	297121.000	1000			
Corrected Total	21793.351	999			

a. R Squared = .564 (Adjusted R Squared = .563)

交互项不显著，将其去掉，进一步分析。Prog 显著，说明参加了政府工作

计划的人的收入明显提高了。提高了多大呢，可以观察系数图。

Parameter Estimates

Dependent Variable:Income after the program

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	4.197	.556	7.548	.000	3.106	5.288
incbef	1.636	.060	27.397	.000	1.519	1.753
[prog=0]	-4.357	.195	-22.290	.000	-4.741	-3.974
[prog=1]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

说明，项目实施前两个收入相同的人，在项目实施后，没有参加政府工作计划的人的收入平均说来要比参加了计划的人的收入低 4.357 元。

Stata 分析:

数据：workprog.dta

命令：anova incaft i.prog c.incbef i.prog#c.incbef

案例五、食物支出的蒙地卡罗实验

注：吴伟改编，对外经济贸易大学，本案例只应用于教学目标

摘要：本案例主要用于理解 OLS 的思想和原理，会建立简单的非线性模型

关键词：线性回归、OLS、蒙地卡罗模拟、Stata

背景描述：应用 Stata 进行数据回归析和 OLS 原理理解

1.教学目的与要求：适用课程:《数据分析与建模》、《公共管理方法与技术》、《社会经济统计学》，使学生理解思想和原理。面向本科经济学、管理学。

2.启发思考题：使用数据自己进行操作分析

3.应用 stata13.1 进行方差分析

4.代码

5.建议学时，4 学时

一、收入与食物支出的关系

* open food data

```
log using food,dta replace text
use food, clear
* examine data
describe
* browse
list
list in 1/5
list food_exp in 1/5
list food_exp if income < 10
* compute summary statistics
summarize
* summarize food expenditure with detail
summarize food_exp, detail
* simple plot data
twoway (scatter food_exp income)
graph save food1, replace // open for editing with: graph use food1
* save graph using saving
twoway (scatter food_exp income), saving(food1, replace)
* store the graph in memory only
twoway (scatter food_exp income), name(food1, replace)
* enhanced plot /* with comments */
twoway (scatter food_exp income), /// /* basic plot control */
ylabel(0(100)600) /// /* Y axis 0 to 600 with ticks each 100 */
xlabel(0(5)35) /// /* X axis 0 to 35 with ticks each 5 */
title(Food Expenditure Data) /* graph title */
graph save food2, replace
* compute least squares regression
regress food_exp income
* calculate fitted values & residuals
predict yhat, xb
```

```
predict ehat, residuals
* compute elasticity at means
margins, eyex(income) atmeans
* compute average of elasticities at each data point
margins, eyex(income)
generate elas = _b[income]*income/yhat
summarize elas
* plot fitted values and data scatter
twoway (scatter food_exp income) /// /* basic plot control */
(lfit food_exp income), /// /* add linear fit */
ylabel(0(100)600) /// /* label Y axis */
xlabel(0(5)35) /// /* label X axis */
title(Fitted Regression Line) /* graph title */
graph save food3, replace
* examine variances and covariances
estat vce
* add observation to data file
edit
set obs 41
replace income=20 in 41
* obtain prediction
predict yhat0
list income yhat0 in 41
log close
* to save changes to food data
* save chap02.dta, replace
* Chapter 2.8.2 Using a Quadratic Model
* new log file
log using chap02_quad, replace text
* open br data and examine
```

```
use br, clear
describe
summarize
* create new variable
generate sqft2=sqft^2
* regression
regress price sqft2
predict priceq, xb
* plot fitted line
twoway (scatter price sqft) /// /* basic plot */
(line priceq sqft, /// /* 2nd plot: line is continuous */
sort lwidth(medthick)) /* sort & change line thickness */
graph save br_quad, replace
* slope and elasticity calculations
di "slope at 2000 = " 2*_b[sqft2]*2000
di "slope at 4000 = " 2*_b[sqft2]*4000
di "slope at 6000 = " 2*_b[sqft2]*6000
di "predicted price at 2000 = " _b[_cons]+_b[sqft2]*2000^2
di "predicted price at 4000 = " _b[_cons]+_b[sqft2]*4000^2
di "predicted price at 6000 = " _b[_cons]+_b[sqft2]*6000^2
di "elasticity at 2000 = " 2*_b[sqft2]*2000^2/(_b[_cons]+_b[sqft2]*2000^2)
di "elasticity at 4000 = " 2*_b[sqft2]*4000^2/(_b[_cons]+_b[sqft2]*4000^2)
di "elasticity at 6000 = " 2*_b[sqft2]*6000^2/(_b[_cons]+_b[sqft2]*6000^2)
* using factor variables
regress price c.sqft#c.sqft
predict price2
margins, dydx(*) at(sqft=(2000 4000 6000))
margins, eyex(*) at(sqft=(2000 4000 6000))
margins, eyex(*)
regress, coeflegend
```

```
generate elas2 = 2*_b[c.sqft#c.sqft]*(sqft^2)/price2
summarize elas2
log close
* Using a Log-linear Model
log using chap02_llin, replace text
use br, clear
* distribution of prices
summarize price, detail
histogram price, percent
graph save price, replace
* distribution of log(price)
generate lprice = ln(price)
histogram lprice, percent
graph save lprice, replace
* log-linear regression
reg lprice sqft
predict lpricef, xb
* price prediction using anti-log
generate pricef = exp(lpricef)
twoway (scatter price sqft) ///
(line pricef sqft, sort lwidth(medthick))
graph save br_loglin, replace
* slope and elasticity calculations
di "slope at 100000 = " _b[sqft]*100000
di "slope at 500000 = " _b[sqft]*500000
di "elasticity at 2000 = " _b[sqft]*2000
di "elasticity at 4000 = " _b[sqft]*4000
* average marginal effects
generate me = _b[sqft]*pricef
summarize me
```

```
generate elas = _b[sqft]*sqft
summarize elas
log close
* Regression with Indicator Variables
* open new log
log using chap02_indicator, replace text
* open utown data and examine
use utown, clear
describe
summarize
* histograms of utown data by neighborhood
histogram price if utown==0, width(12) start(130) percent ///
xtitle(House prices ($1000) in Golden Oaks) ///
xlabel(130(24)350) legend(off)
graph save utown_0, replace
histogram price if utown==1, width(12) start(130) percent ///
xtitle(House prices ($1000) in University Town) ///
xlabel(130(24)350) legend(off)
graph save utown_1, replace
graph combine "utown_0" "utown_1", col(1) iscale(1)
graph save combined, replace
* using by option
label define utownlabel 0 "Golden Oaks" 1 "University Town"
label value utown utownlabel
histogram price, by(utown, cols(1)) ///
start(130) percent ///
xtitle(House prices ($1000)) ///
xlabel(130(24)350) legend(off)
graph save combined2, replace
* summary stats
```

```
summarize price if utown==0
summarize price if utown==1
* summary stats using by
by utown, sort: summarize price
* summary stats using bysort
bysort utown: summarize price
* regression
regress price utown
* test of two means
ttest price, by(utown)
log close
* calculation of Average marginal effects
* food expenditure example
log using food_me, replace text
use food, clear
summarize income
return list
scalar xbar = r(mean)
quietly regress food_exp income
margins, eyex(*) atmeans
nlcom _b[income]*xbar/(_b[_cons]+_b[income]*xbar)
log close
* quadratic house price example
log using chap02_quad_me, replace text
use br, clear
quietly regress price c.sqft#c.sqft
margins, eyex(*) at(sqft=2000)
nlcom 2*_b[c.sqft#c.sqft]*(2000^2)/(_b[_cons]+_b[c.sqft#c.sqft]*(2000^2))
log close
* slope in log-linear model
```

```
log using chap02_llin_me, replace text
use br, clear
gen lprice = log(price)
quietly regress lprice sqft
nlcom _b[sqft]*exp(_b[_cons]+_b[sqft]*2000)
log close
```

二、蒙地卡罗模拟实验

```
*clear memory and start new log
clear all
log using chap02_app2G, replace text
* define some global macros
global numobs 40 // sample size
global beta1 100 // intercept parameter
global beta2 10 // slope parameter
global sigma 50 // error standard deviation
* random number seed
set seed 1234567
* create artificial data using y = beta1+beta2*x+e
set obs $numobs
generate x = 10
replace x = 20 if _n > $numobs/2
generate y = $beta1 + $beta2*x + rnormal(0,$sigma)
* regression with artifical data
regress y x
di "rmse " e(rmse)
estat vce
* data file mc1.data created using following command
save mc1, replace
* program to generate data and estimate regression
program chap02sim, rclass
```

```
version 11.1
drop _all
set obs $numobs
generate x = 10
replace x = 20 if _n > $numobs/2
generate ey = $beta1 + $beta2*x
generate e = rnormal(0, $sigma)
generate y = ey + e
regress y x
return scalar b2 =_b[x] // saves slope
return scalar b1 =_b[_cons] // saves intercept
return scalar sig2 = (e(rmse))^2 // saves sigma^2
end
* simulate command
simulate b1r = r(b1) b2r=r(b2) sig2r=r(sig2) , ///
reps(1000) nodots nolegend seed(1234567): chap02sim
* display experiment parameters
di " Simulation parameters"
di " beta1 = " $beta1
di " beta2 = " $beta2
di " N = " $numobs
di " sigma^2 = " $sigma^2
* summarize experiment results
summarize, detail
* histogram sampling distribution of LS estimates
histogram b2r, percent normal
graph save b2r, replace
log close
```

案例六、新垃圾焚烧炉的建设对住房价值的影响

注：吴伟改编，对外经济贸易大学，本案例只应用于教学目标

摘要：本案例主要用于理解 DID 的思想和原理，会建立简单的 DID 模型

关键词：两期面板、DID、Stata

背景描述：应用 Stata 进行 DID 分析。应用双重差分评价公共项目的经济影响

1.教学目的与要求：适用课程:《数据分析与建模》、《公共管理方法与技术》、《社会经济统计学》，使学生理解思想和原理。面向本科经济学、管理学。

2.启发思考题：使用数据自己进行操作分析

3.应用 stata13.1 进行方差分析

4.代码

5.建议学时，4 学时

数据：KIELMC.dta

假设是相对于较远的房子，邻近焚烧炉的住房价格将会下降。

$$\widehat{rprice} = \underset{(3,093.0)}{101,307.5} - \underset{(5,827.71)}{30,688.27}\ nearinc$$

$$\widehat{rprice} = \underset{(2,653.79)}{82,517.23} - \underset{(4,744.59)}{18,824.37}\ nearinc$$

如果根据系数，认为修建焚烧厂后，对于住房价格产生了极强的负向影响就错了。

我们需要与修建前的情况进行比较：

$\hat{\delta}_1 = -30,688.27 - (-18,824.37) = -11,863.9$

说明没有那么大的负向影响

这等同于：

$\hat{\delta}_1 = (\overline{rprice}_{1,nr} - \overline{rprice}_{1,fr}) - (\overline{rprice}_{0,nr} - \overline{rprice}_{0,fr})$

这种方法，被称为双重差分法(difference-in-differences estimator (DiD)

$rprice = \beta_0 + \delta_0\, after + \beta_1\, nearinc + \delta_1\, after{\cdot}nearinc + u$

通过这种方式，双重差分效应的标准误可以得到

如果修建焚烧炉前后销售的住房存在系统性差异，其他的解释变量需要引入

（协变量/控制变量）

这种模型也可以减少误差方差乃至标准误

Stata 代码

codebook, compact

reg rprice nearinc if year==1981

```
. reg rprice nearinc if year==1981
```

Source	SS	df	MS	Number of obs =	142
Model	2.7059e+10	1	2.7059e+10	F(1, 140) =	27.73
Residual	1.3661e+11	140	975815048	Prob > F =	0.0000
				R-squared =	0.1653
Total	1.6367e+11	141	1.1608e+09	Adj R-squared =	0.1594
				Root MSE =	31238

rprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
nearinc	-30688.27	5827.709	-5.27	0.000	-42209.97	-19166.58
_cons	101307.5	3093.027	32.75	0.000	95192.43	107422.6

reg rprice nearinc if year==1978

```
. reg rprice nearinc if year==1978
```

Source	SS	df	MS	Number of obs =	179
Model	1.3636e+10	1	1.3636e+10	F(1, 177) =	15.74
Residual	1.5332e+11	177	866239953	Prob > F =	0.0001
				R-squared =	0.0817
Total	1.6696e+11	178	937979126	Adj R-squared =	0.0765
				Root MSE =	29432

rprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
nearinc	-18824.37	4744.594	-3.97	0.000	-28187.62	-9461.117
_cons	82517.23	2653.79	31.09	0.000	77280.09	87754.37

$$\widehat{rprice} = \underset{(3,093.0)}{101,307.5} - \underset{(5,827.71)}{30,688.27}\ nearinc$$

$$\widehat{rprice} = \underset{(2,653.79)}{82,517.23} - \underset{(4,744.59)}{18,824.37}\ nearinc$$

使用双重差分模型一次性计算得出

reg rprice y81 nearinc y81#nearinc

Source	SS	df	MS	Number of obs =	321
				F(3, 317) =	22.25
Model	6.1055e+10	3	2.0352e+10	Prob > F =	0.0000
Residual	2.8994e+11	317	914632739	R-squared =	0.1739
				Adj R-squared =	0.1661
Total	3.5099e+11	320	1.0969e+09	Root MSE =	30243

rprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
y81	18790.29	4050.065	4.64	0.000	10821.88	26758.69
nearinc	-30688.27	5642.057	-5.44	0.000	-41788.88	-19587.66
y81#nearinc						
0 1	11863.9	7456.646	1.59	0.113	-2806.867	26534.67
1 0	0	(omitted)				
1 1	0	(omitted)				
_cons	82517.23	2726.91	30.26	0.000	77152.1	87882.36

案例七、啤酒税与交通事故死亡率

注：吴伟改编，对外经济贸易大学，本案例只应用于教学目标

摘要：本案例主要用于理解面板数据的思想和原理，会建立简单的固定效应及随机效应模型

关键词：两期面板、DID、Stata

背景描述：应用 Stata 进行面板数据分析。应用面板数据评价公共项目的经济影响

1.教学目的与要求：适用课程:《数据分析与建模》、《公共管理方法与技术》、《社会经济统计学》，使学生理解思想和原理。面向本科公共管理学生。

2.启发思考题：使用数据自己进行操作分析

3.应用 stata13.1 进行方差分析

4.代码

5.建议学时，6 学时

数据：fatality.dta

变量定义：mrall:交通事故死亡率，beertax:啤酒税

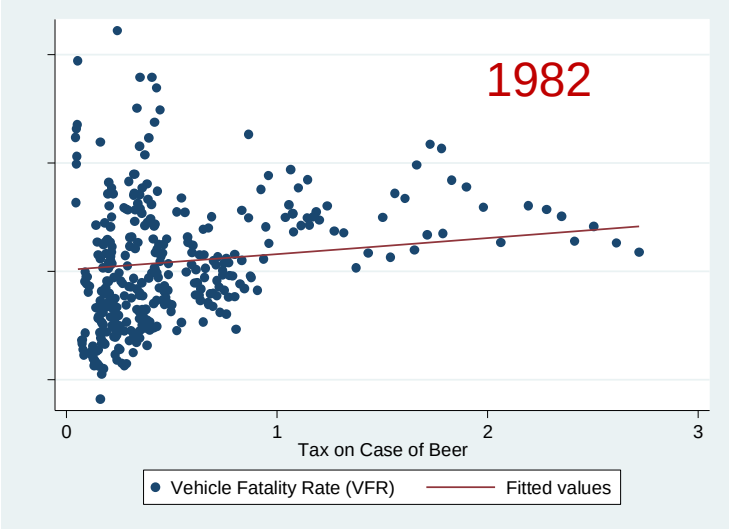
先做 1982 年和 1988 年的散点图

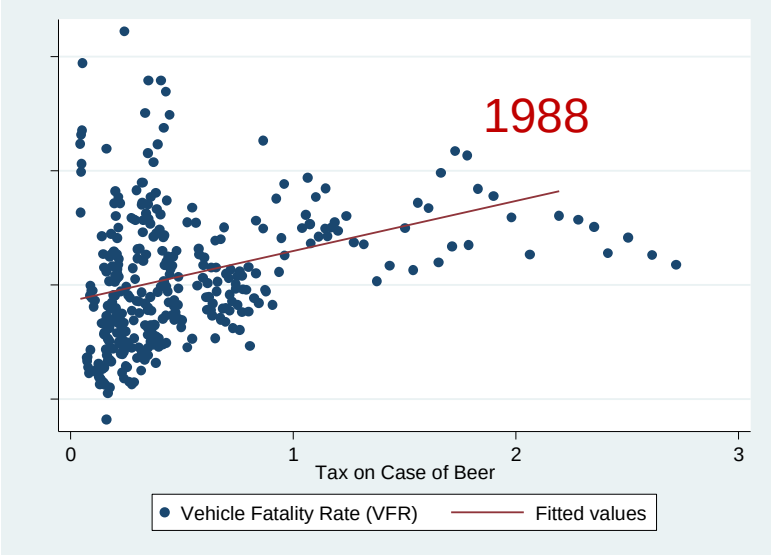
scatter mrall beertax || lfit mrall beertax if year==1982

scatter mrall beertax || lfit mrall beertax if year==1988

做啤酒税和死亡率的 OLS

reg mrall beertax if year==1988





啤酒税越高，死亡率越高，与日常推理不相符？

为什么？存在内生性问题，具有未关测到的效应与解释变量相关。应用面板数据可以最大程度减少内生性问题。

不能就此得出结论啤酒税无助于降低死亡率。

事实上，有很多省略变量如道路状况、驾驶环境（城市/农村）、汽车密度、关于酒后驾车的文化，都会可能与啤酒税相关，故而这时具有省略变量导致的内生性问题。

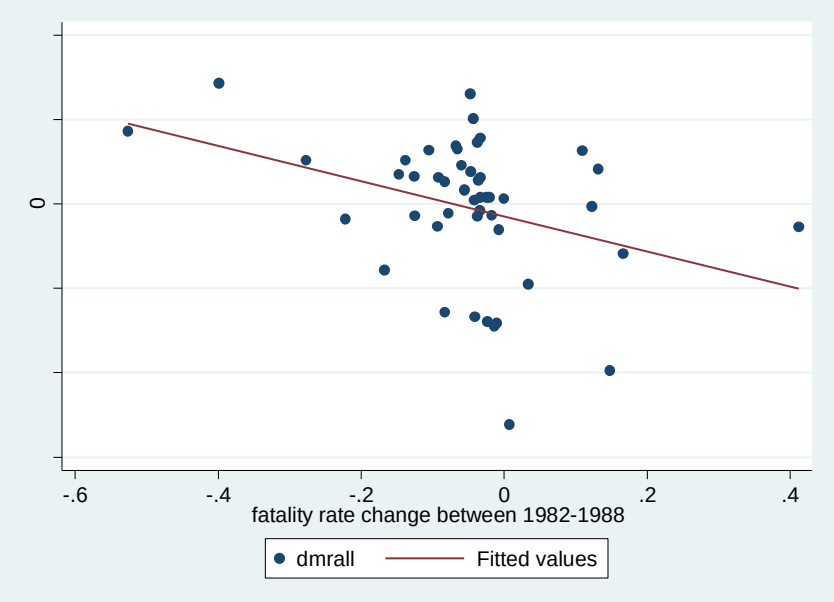
一种办法就是收集这些变量的数据，把它加入截面回归的解释变量中，然而，饮酒和酒驾的文化难以度量。如果这些因素对于给定的州，不随时间而变，那么我们可以用面板数据的固定效应模型加以解决，尽管我们不知道这些变量的值。

直觉上，某一州对酒驾的文化态度影响饮酒水平以及交通死亡率，如果他们在 1982-1988 年间不发生变化，那么他们也不会改变死亡率。

伴随时间变化的死亡率的变化来自于其他因素，上述方程中，就是啤酒税的变化和误差项的变化。他们捕获了影响死亡率的其他因素的变化。

- $FatalityRate_{i1988} - FatalityRate_{i1982} = \beta_1(BeerTax_{i1988} - BeerTax_{i1982}) + u_{i1988} - u_{i1982}$

```
keep if year==1982 | year==1988
codebook, compact
gen dmrall=mrall-mrall[_n-1] if year==year[_n-1]+6
list mrall dmrall
gen dbeertax = beertax - beertax[ _n-1] if year==year[_n-1]+6
scatter dmrall dbeertax
reg dmrall dbeertax
```



现在，啤酒税显著低降低了交通事故死亡率。通过了 0.05（5%）的显著性水平检验。

啤酒税每增加 1 美元每箱，将减少死亡率 1.04/每万人，考虑到每年每万人的平均死亡率约为 2,提高 1 美元的税，可以降低一半死亡率。

```
. reg dmrall dbeertax
```

Source	SS	df	MS	Number of obs	=	48
				F(1, 46)	=	6.22
Model	9.6645e-09	1	9.6645e-09	Prob > F	=	0.0162
Residual	7.1418e-08	46	1.5526e-09	R-squared	=	0.1192
				Adj R-squared	=	0.1000
Total	8.1082e-08	47	1.7251e-09	Root MSE	=	3.9e-05

dmrall	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
dbeertax	-.0001041	.0000417	-2.49	0.016	-.0001881	-.0000201
_cons	-7.20e-06	6.06e-06	-1.19	0.241	-.0000194	5.00e-06

一次差分考虑了不随时间变化的固定因素，但是依然可能存在随着时间变化的因素(time varying factors)，他们也可能影响交通事故死亡率，同时与啤酒税具有相关性，省略了这些随时间变化因素将导致省略变量偏误。

此外，当我们的数据时期 T>2，我们只用两期的数据非常愚蠢，我们希望利用所有的数据进行估计，这样，我们使用面板数据的固定效应模型(fixed effect model)

xtset state year

xtreg mrall beertax, fe

```
. xtreg mrall beertax, fe

Fixed-effects (within) regression               Number of obs      =       336
Group variable: state                           Number of groups   =        48

R-sq:  within  = 0.0407                         Obs per group: min =         7
       between = 0.1101                                        avg =       7.0
       overall = 0.0934                                        max =         7

                                                F(1,287)           =     12.19
corr(u_i, Xb)  = -0.6885                        Prob > F           =    0.0006
```

mrall	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
beertax	-.0000656	.0000188	-3.49	0.001	-.0001026	-.0000286
_cons	.0002377	9.70e-06	24.51	0.000	.0002186	.0002568
sigma_u	.00007147					
sigma_e	.00001899					
rho	.93408484	(fraction of variance due to u_i)				

```
F test that all u_i=0:     F(47, 287) =    52.18             Prob > F = 0.0000
```

由于包含更多期，啤酒税系数小于一次差分的估计，但是依然显著。由于增加了样本，标准误变小。固定效用模型使我们消除了不随时间变化的省略变量。

然而，依然还有其他一些因素导致省略变量偏误，例如，随着时间变化汽车越来越安全，乘员习惯系安全带，啤酒税率随着时间推移提高，这样，提高了总体汽车安全效果，这些安全改进随时间演化，但是对于所有州相同，这样，我们可以引入时间固定效应(time fixed effect)消除影响。

```
. xtreg mrall beertax i.year, fe

Fixed-effects (within) regression               Number of obs      =       336
Group variable: state                           Number of groups   =        48

R-sq:  within  = 0.0803                         Obs per group: min =         7
       between = 0.1101                                        avg =       7.0
       overall = 0.0876                                        max =         7

                                                F(7,281)           =      3.50
corr(u_i, Xb)  = -0.6781                        Prob > F           =    0.0013
```

mrall	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
beertax	-.000064	.0000197	-3.24	0.001	-.0001029	-.0000251
year						
1983	-7.99e-06	3.84e-06	-2.08	0.038	-.0000155	-4.41e-07
1984	-7.24e-06	3.84e-06	-1.89	0.060	-.0000148	3.07e-07
1985	-.0000124	3.84e-06	-3.23	0.001	-.00002	-4.83e-06
1986	-3.79e-06	3.86e-06	-0.98	0.327	-.0000114	3.81e-06
1987	-5.09e-06	3.90e-06	-1.31	0.193	-.0000128	2.58e-06
1988	-5.18e-06	3.96e-06	-1.31	0.192	-.000013	2.62e-06
_cons	.0002428	.0000108	22.46	0.000	.0002216	.0002641
sigma_u	.00007095					
sigma_e	.00001879					
rho	.93446372	(fraction of variance due to u_i)				

```
F test that all u_i=0:     F(47, 281) =    53.19             Prob > F = 0.0000
```

到现在为止，我们的实证完美吗？

也许不，酒精税也许只是抑制交通事故死亡率的一种方法，还有相关法律可能起作用，如果忽略他们，会导致省略变量偏误。如最低饮酒年龄、强制蹲监狱、社区矫正等。

还有其他社会经济变量如驾驶员的技术、失业率、人均收入等等也会影响死亡率。

当我们把所有的因素都找到了，并且有数据，我们或许可以研究交通事故死亡率的影响因素了，而不是啤酒税或交通法对死亡率的影响了！

考虑所有可能的解释变量，并纠正异方差问题。

```
gen age=int(mlda)
gen lperinc=log(perinc)
xtreg mrall beertax i.year unrate lperinc jaild comserd vmiles ib21.age , fe
```

xtreg mrall beertax i.year unrate lperinc jaild comserd vmiles ib21.age , fe vce(cluster state)

mrall	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
beertax	-.0000449	.0000298	-1.50	0.139	-.0001049	.0000151
year						
1983	-9.81e-06	3.10e-06	-3.16	0.003	-.000016	-3.57e-06
1984	-.0000287	4.76e-06	-6.03	0.000	-.0000383	-.0000192
1985	-.0000379	5.19e-06	-7.30	0.000	-.0000483	-.0000274
1986	-.0000344	6.52e-06	-5.28	0.000	-.0000475	-.0000213
1987	-.0000446	8.01e-06	-5.57	0.000	-.0000607	-.0000285
1988	-.0000533	9.23e-06	-5.78	0.000	-.0000719	-.0000347
unrate	-6.31e-06	1.30e-06	-4.84	0.000	-8.94e-06	-3.69e-06
lperinc	.0001815	.0000638	2.85	0.007	.0000532	.0003097
jaild	1.31e-06	1.67e-06	0.78	0.437	-2.05e-06	4.66e-06
comserd	3.34e-06	.0000133	0.25	0.803	-.0000234	.0000301
vmiles	8.23e-10	6.86e-10	1.20	0.236	-5.57e-10	2.20e-09
age						
18	2.81e-06	7.03e-06	0.40	0.691	-.0000113	.000017
19	-1.86e-06	4.98e-06	-0.37	0.711	-.0000119	8.16e-06
20	3.14e-06	5.04e-06	0.62	0.537	-7.00e-06	.0000133
_cons	-.0014325	.0006128	-2.34	0.024	-.0026653	-.0001998
sigma_u	.00009023					
sigma_e	.00001556					
rho	.97113027	(fraction of variance due to u_i)				

几点有意思的发现：

增加了解释变量，降低了啤酒税的系数(-0.000045)，然而系数仍然显著(0.05)

如果某个州打算提高平均啤酒税一倍，由于目前的平均税为 0.5,则再提高 0.5/每箱。这样，啤酒税每增加 0.5 美元，万人死亡率降低 0.45*0.5=0.23 人，这一效应很大，因为每万人死亡率为 2，相当于减低万人死亡率至 1.77 人。标准误为 0.22，比较大，表明这种预测不够精确。

最低合法饮酒年龄效果低，且不太显著。投入监狱也不显著。

相关社会经济变量通过了显著性检验，失业率越高，死亡率越低。失业率增长一个百分点，死亡率下降 0.063/万人；人均收入也显著，收入增长 1%，死亡率上升 0.0182/每万人。

良好的经济条件与高死亡率相关，也许是因为失业率降低了交通密度；收入提高使人们的酒精消费量提高，故而增加了交通死亡率。

pwcorr spircons perinc, sig

发现人均收入提高与酒精消费正相关，很显著。