



电子科技大学 经济与管理学院
School of Management and Economics of UESTC

计量经济学

Econometrics

任课老师：李亚静

电子科大经管学院



第五讲 虚拟变量

(教材3.3、 7.4)

第八讲 虚拟变量



什么是虚拟变量?

❖ 经济变量

- **可以定量度量**: 商品需求量、价格、收入、产量等
- **无法定量度量**: 职业、性别、战争、自然灾害等

❖ 虚拟变量 (dummy variables) : 定性变量

第八讲 虚拟变量



引例

- ❖ 比较某个国家三个地区（东部、中部和西部）公立学校教师平均薪水在统计上的差异
 - 1，各个地区之间是否存在差异？统计上怎样检验？
 - 2，如果存在差异，**地理位置**是关键因素吗？

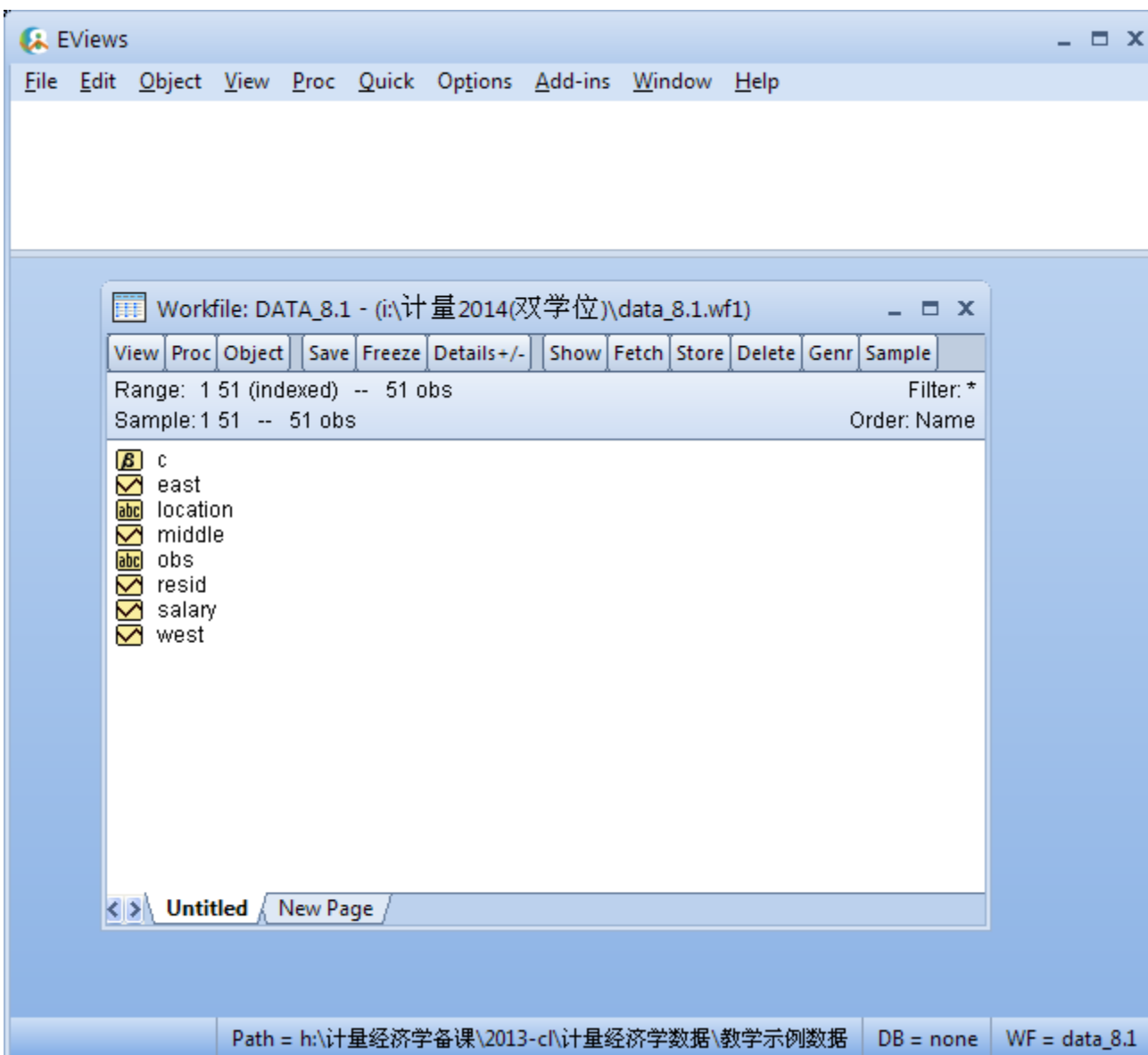
第八讲 虚拟变量



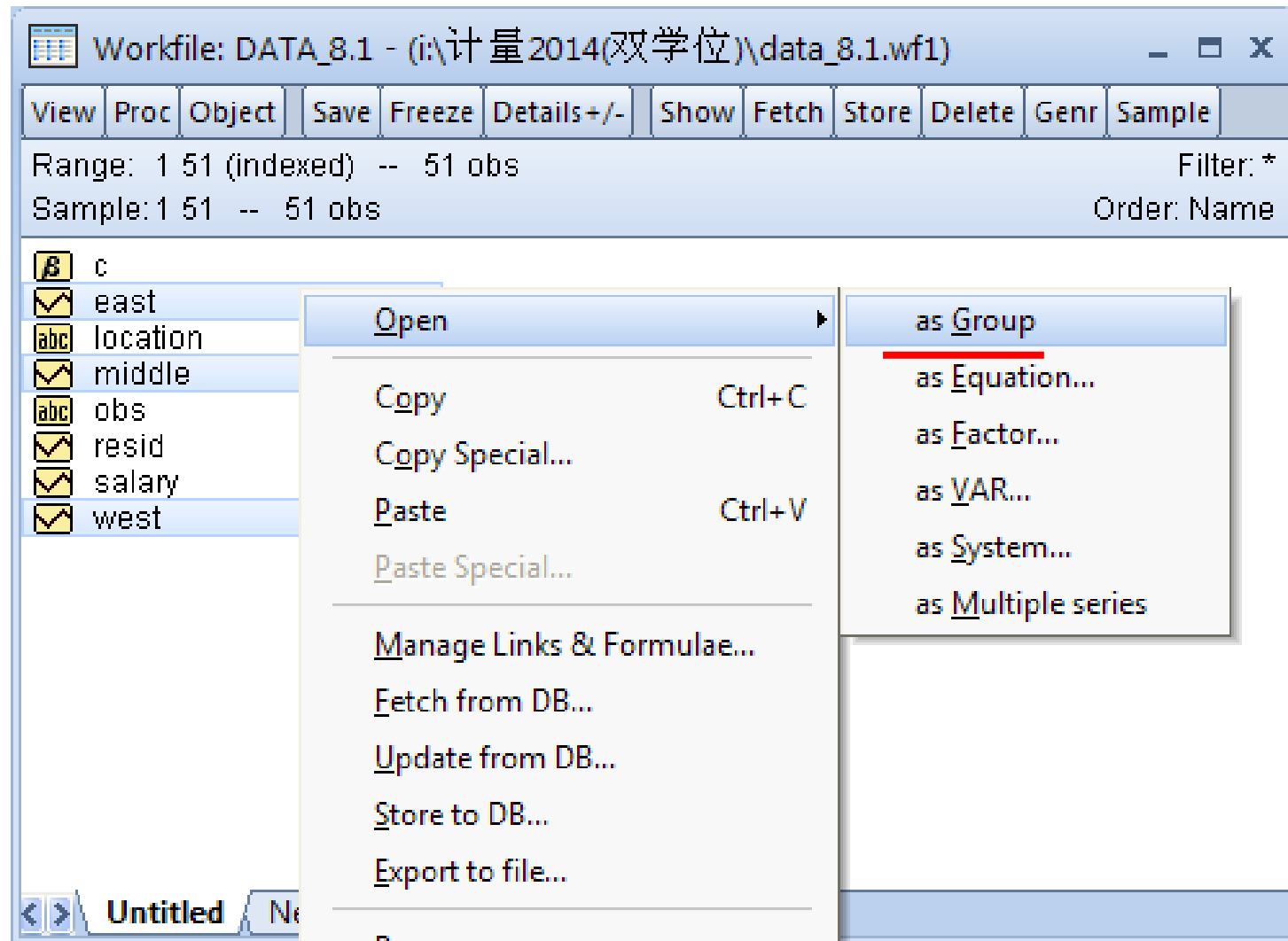
引例

- ❖ 从51个县（市）收集了当地教师的平均年薪数据，并处理成Excel文档（**data_8.1**）。
- 东部：13
- 中部：21
- 西部：17

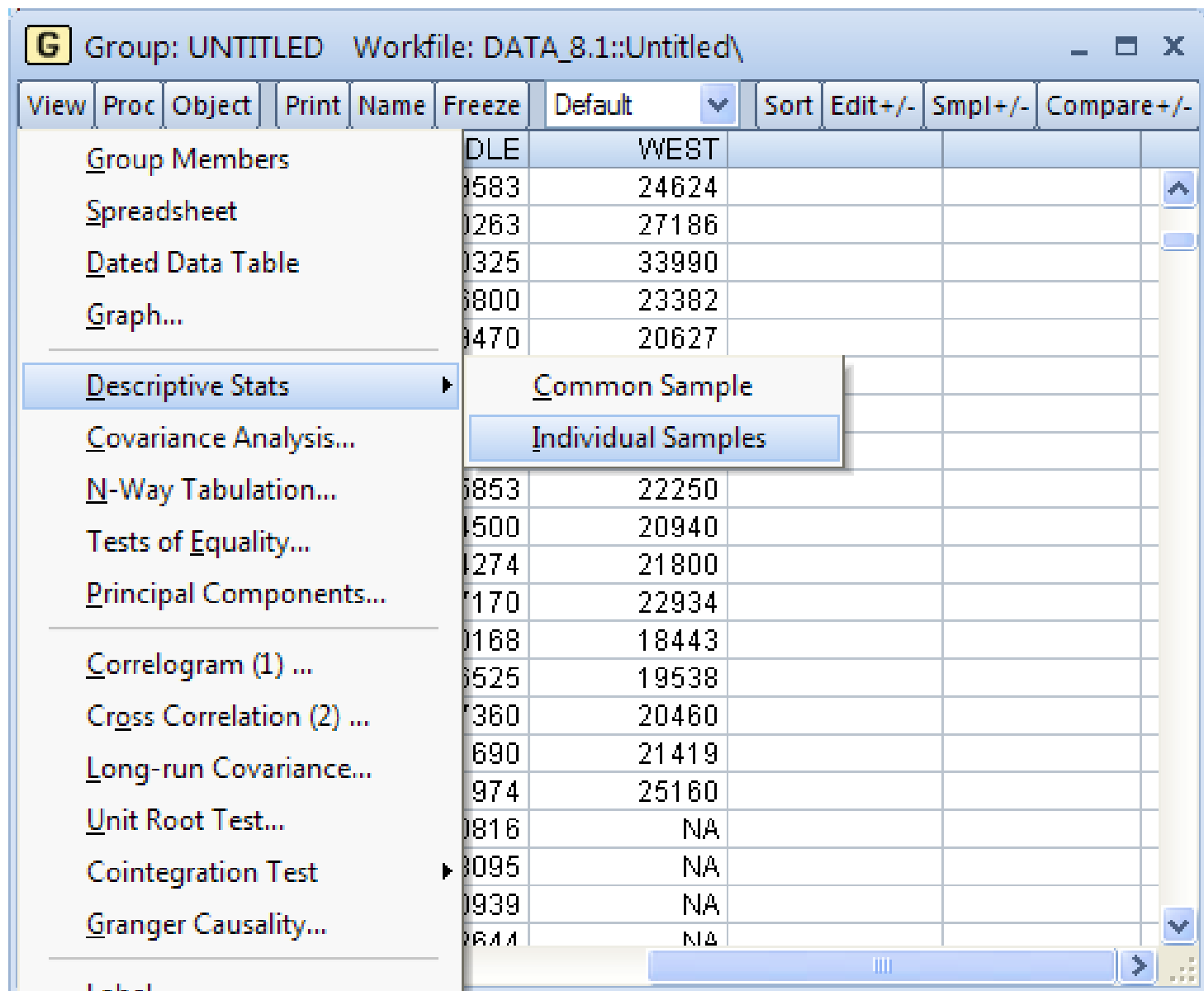
	A	B	C	D	E	F	G	
1	OBS	LOCATION	SALARY	EAST	MIDDLE	WEST		
2	1	中部	19583	22482	19583	24624		
3	2	中部	20263	20969	20263	27186		
4	3	中部	20325	27224	20325	33990		
5	4	中部	26800	25892	26800	23382		
6	5	中部	29470	22644	29470	20627		
7	6	中部	26610	24640	26610	22795		
8	7	中部	30678	22341	30678	21570		
9	8	中部	27170	25610	27170	22080		
10	9	中部	25853	26015	25853	22250		
11	10	中部	24500	25788	24500	20940		
12	11	中部	24274	29132	24274	21800		
13	12	中部	27170	41480	27170	22934		
14	13	中部	30168	25845	30168	18443		
15	14	中部	26525		26525	19538		
16	15	中部	27360		27360	20460		
17	16	中部	21690		21690	21419		
18	17	中部	21974		21974	25160		
19	18	中部	20816		20816			
20	19	中部	18095		18095			
21	20	中部	20939		20939			
22	21	中部	22644		22644			
23	22	西部	24624					
24	23	西部	27186					



描述性统计分析



Group: UNTITLED Workfile: DATA_8.1::Untitled\						
View	Proc	Object	Print	Name	Freeze	Default
		EAST	MIDDLE	WEST		
1		22482	19583	24624		
2		20969	20263	27186		
3		27224	20325	33990		
4		25892	26800	23382		
5		22644	29470	20627		
6		24640	26610	22795		
7		22341	30678	21570		
8		25610	27170	22080		
9		26015	25853	22250		
10		25788	24500	20940		
11		29132	24274	21800		
12		41480	27170	22934		
13		25845	30168	18443		
14		NA	26525	19538		
15		NA	27360	20460		
16		NA	21690	21419		
17		NA	21974	25160		
18		NA	20816	NA		
19		NA	18095	NA		
20		NA	20939	NA		
21		NA	22644	NA		
22						



G Group: UNTITLED Workfile: DATA_8.1::Untitled\									
View	Proc	Object	Print	Name	Freeze	Sample	Sheet	Stats	Spec
				EAST	MIDDLE	WEST			
Mean				26158.62	24424.14	22894.00			
Median				25788.00	24500.00	22080.00			
Maximum				41480.00	30678.00	33990.00			
Minimum				20969.00	18095.00	18443.00			
Std. Dev.				5123.734	3725.544	3553.857			
Skewness				2.148434	0.027051	1.858708			
Kurtosis				7.396038	1.822553	6.718643			
Jarque-Bera				20.46862	1.215644	19.58364			
Prok									
Sum				340062.0	512907.0	389198.0			
Sum Sq. Dev.				3.15E+08	2.78E+08	2.02E+08			
Observations				13	21	17			

仅从均值来看，你的结论是什么？

怎样检验东部和西部在均值是否存在显著差异？

Workfile: DATA_8.1 - (i:\计量2014(双学位)\data_8.1.wf1)

View Proc Object Save Freeze Details+/- Show Fetch Store Delete Genr Sample

Range: 1 51 (indexed) -- 51 obs Filter: *
Sample: 1 51 -- 51 obs Order: Name

Group: UNTITLED Workfile: DATA_8.1::Untitl...

	EAST	WEST
1	22482	24624
2	20969	27186
3	27224	33990
4	25892	23382
5	22644	20627
6	24640	22795
7	22341	21570
8	25610	22080
9	26015	22250
10	25788	20940
11	29132	21800
12	41480	22934
13	25845	18443
14	NA	19538
15	NA	20460
16	NA	21419
17	NA	25160
18	NA	NA
19		

Group: UNTITLED Workfile: DATA_8.1::Untitl... - □ X

View	Proc	Object	Print	Name	Freeze	Default	Sort	Edit
Group Members						EST		
Spreadsheet						1624		^
Dated Data Table						1186		
Graph...						8990		
						8382		
						0627		
						0705		
Descriptive Stats								
Covariance Analysis...								
N-Way Tabulation...								
Tests of Equality...								
Principal Components...								
Correlogram (1) ...								
Cross Correlation (2) ...								
Long-run Covariance...								
Unit Root Test...								
Cointegration Test								

Test Between Series X

Test equality of

☒ Mean

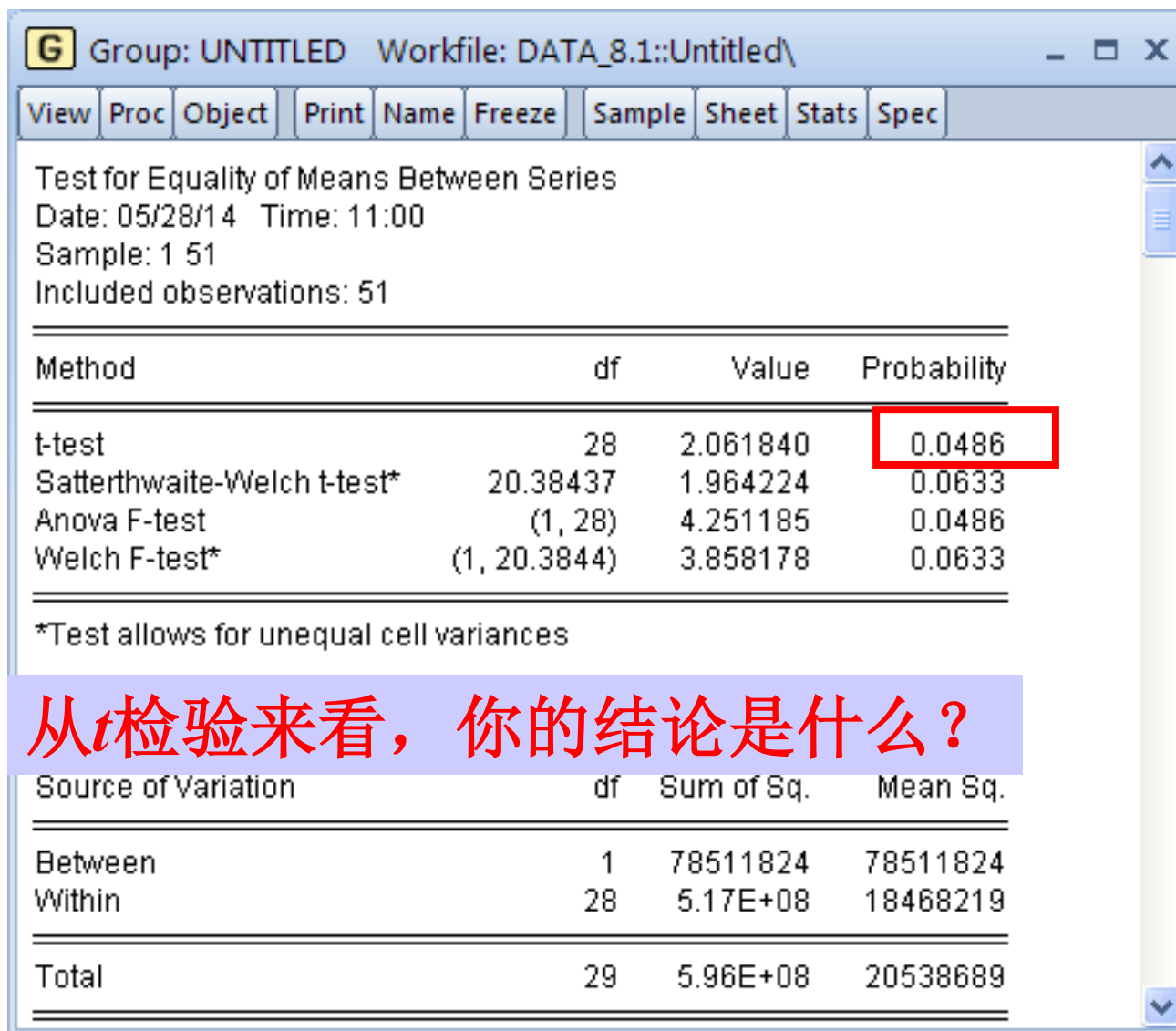
☐ Median

☐ Variance

☐ Common sample

OK

Cancel



第八讲 虚拟变量

思考?

- ❖ 地理位置真的是影响教师薪水差异的关键因素吗?
- ❖ 除了地理位置，还存在其他因素吗?
- ❖ 怎样排除其他因素的影响?



第八讲 虚拟变量

引例：虚拟变量模型

- ❖ 比较某个国家三个地区（东部、中部和西部）公立学校教师平均薪水在统计上的差异(**data_8.2**)

$$Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \mu_i$$

Y_i = 第 i 个县的平均薪水

为什么三个地区只设两个虚拟变量？

$$D_{1i} = \begin{cases} 1 & \text{若该县位于中部} \\ 0 & \text{其它} \end{cases}$$

$$D_{2i} = \begin{cases} 1 & \text{若该县位于西部} \\ 0 & \text{其它} \end{cases}$$

第八讲 虚拟变量



引例：虚拟变量模型

给定： $Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \mu_i$

中部平均薪水：

$$E(Y_i | D_{1i} = 1, D_{2i} = 0) = \beta_0 + \beta_1$$

西部平均薪水：

$$E(Y_i | D_{1i} = 0, D_{2i} = 1) = \beta_0 + \beta_2$$

东部平均薪水：

$$E(Y_i | D_{1i} = 0, D_{2i} = 0) = \beta_0$$

	SALARY	LOCATION	SPENDING	D1	D2
1	19583	中部	3346	1	0
2	20263	中部	3114	1	0
3	20325	中部	3554	1	0
4	26800	中部	4642	1	0
5	29470	中部	4669	1	0
6	26610	中部	4888	1	0
7	30678	中部	5710	1	0
8	27170	中部	5536	1	0
9	25853	中部	4168	1	0
10	24500	中部	3547	1	0
11	24274	中部	3159	1	0
12	27170	中部	3621	1	0
13	30168	中部	3782	1	0
14	26525	中部	4247	1	0
15	27360	中部	3982	1	0
16	21690	中部	3568	1	0
17	21974	中部	3155	1	0
18	20816	中部	3059	1	0
19	18095	中部	2967	1	0
20	20939	中部	3285	1	0
21	22644	中部	3914	1	0
22	24624	西部	4517	0	1
23	27186	西部	4349	0	1
24	33990	西部	5020	0	1
25	23382	西部	3594	0	1
26	20627	西部	2821	0	1
27	22795	西部	3366	0	1
28	21570	西部	2920	0	1
29	22080	西部	2980	0	1

D1:中部
D2:西部

Equation Estimation

Specification Options

Equation specification:

Dependent variable followed by list of regressors and PDL terms, OR an explicit equation like

salary c d1 d2

Estimation settings:

Method: LS - Least Squares (NLS and ARMA)

Sample: 1 51

确定 取消

Sample: 1 51
Included observations: 51

D1:中部
D2:西部

Variable	Coefficient	Std. Error	t-Statistic	Prob.
从回归结果来看，你的结论是什么？				.0000
D2	-3264.615	1499.155	-2.177637	.0344
根据回归结果，你有什么建议？				24356.22
Adjusted R-squared	0.052170	S.D. dependent var		4179.426

第八讲 虚拟变量

仅仅是地区因素吗？

根据回归结果，你的结论是什么？

Sample: 1 51

Included observations: 51

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	13269.11	1395.056	9.511530	0.0000
D1	-1673.514	801.1703	-2.088837	0.0422
D2	-1144.157	861.1182	-1.328687	0.1904
<u>SPENDING</u>	3.288848	0.317642	10.35393	0.0000

D1:中部
D2:西部

R-squared	0.722665	Mean dependent var	24356.22
Adjusted R-squared	0.704963	S.D. dependent var	4179.426
S.E. of regression	2270.152	Akaike info criterion	18.36827
Sum of squared resid	21451.88	Schwarz criterion	18.51878

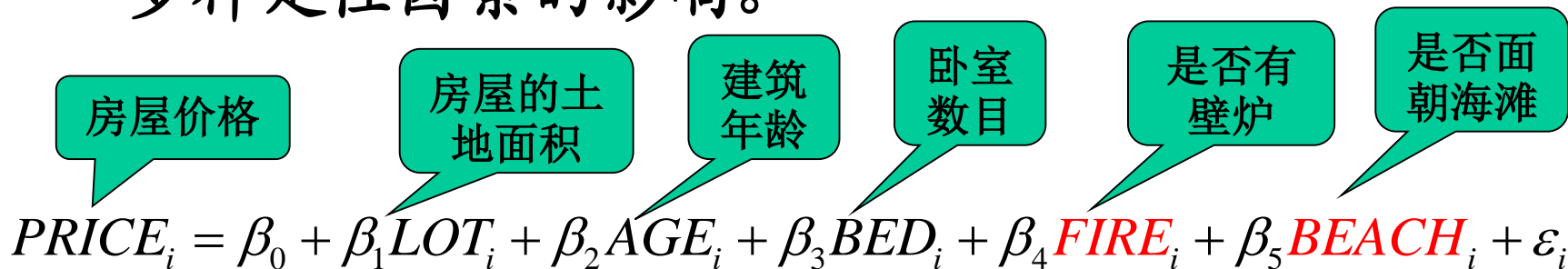
根据新的回归结果，政府应该补贴哪个地区的教师？

Prob(F-statistic) 0.000000

第八讲 虚拟变量

虚拟变量的设置原则

- ❖ 每一定性变量所需的虚拟变量个数要比该定性变量的类别数少1，即如果有 m 个定性变量，只在模型中引入 $m-1$ 个虚拟变量。
- ❖ 在同一个方程中，可以引入多个虚拟变量来考察多种定性因素的影响。



The diagram illustrates a regression model for house prices. The equation is $PRICE_i = \beta_0 + \beta_1 LOT_i + \beta_2 AGE_i + \beta_3 BED_i + \beta_4 FIRE_i + \beta_5 BEACH_i + \varepsilon_i$. Each variable in the equation is linked by a callout bubble to its description: $PRICE_i$ (房屋价格), LOT_i (房屋的土地面积), AGE_i (建筑年龄), BED_i (卧室数目), $FIRE_i$ (是否有壁炉), and $BEACH_i$ (是否面朝海滩). The variables $FIRE_i$ and $BEACH_i$ are highlighted in red in the original image, indicating they are dummy variables.

$$PRICE_i = \beta_0 + \beta_1 LOT_i + \beta_2 AGE_i + \beta_3 BED_i + \beta_4 FIRE_i + \beta_5 BEACH_i + \varepsilon_i$$

第八讲 虚拟变量

虚拟变量的引入

❖ 加法形式：考察截距的不同

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \mu_i$$

单独采用乘法形式的情况极少

❖ 乘法形式：考察斜率的不同

$$C_t = \beta_0 + \beta_1 X_t + \beta_2 D_t X_t + \mu_t$$

$D_i X_i$ ：被称为斜率虚拟变量

❖ 混合形式：截距和斜率同时发生变化

$$Y_i = \beta_0 + \beta_1 X_i + \beta_3 D_i + \beta_4 (D_i X_i) + \mu_i$$

例：考察1992年前后的中国居民的总储蓄-收入关系是否已发生变化。以 Y 为储蓄， X 为收入。

$$\text{1992年前: } Y_i = \alpha_1 + \alpha_2 X_i + \mu_{1i} \quad i=1, 2, \dots, n_1$$

$$\text{1992年后: } Y_i = \beta_1 + \beta_2 X_i + \mu_{2i} \quad i=1, 2, \dots, n_2$$

则有可能出现下述四种情况中的一种：

1. $\alpha_1 = \beta_1$ ，且 $\alpha_2 = \beta_2$ ，称为**重合回归**。
2. $\alpha_1 \neq \beta_1$ ，但 $\alpha_2 = \beta_2$ ，差异仅在其截距，称为**平行回归**。
3. $\alpha_1 = \beta_1$ ，但 $\alpha_2 \neq \beta_2$ ，差异仅在其斜率，称为**同截距回归**。
4. $\alpha_1 \neq \beta_1$ ，且 $\alpha_2 \neq \beta_2$ ，两个回归完全不同，称为**非相似回归**。

将 n_1 与 n_2 次观察值合并，估计以下回归方程：

$$Y_i = \beta_0 + \beta_1 X_i + \beta_3 D_i + \beta_4 (D_i X_i) + \mu_i$$

D_i 为引入的虚拟变量：
$$D_i = \begin{cases} 1 & \text{1992年前} \\ 0 & \text{1992年后} \end{cases}$$

于是有：

$$E(Y_i | D_i = 0, X_i) = \beta_0 + \beta_1 X_i$$

$$E(Y_i | D_i = 1, X_i) = (\beta_0 + \beta_3) + (\beta_1 + \beta_4) X_i$$

可分别表示1992年后面年度和前面年度的储蓄函数。

具体的回归结果为：

$$\hat{Y}_i = -15452 + 0.8881X_i + 13802.3D_i - 0.4765D_iX_i$$

t值 (-6.11) (22.89) (4.33) (-2.55)

$$\bar{R}^2 = 0.9836$$

储蓄函数分别为：

1992年前： $\hat{Y}_i = -1649.7 + 0.4116X_i$

1992年后： $\hat{Y}_i = -15452 + 0.8881X_i$



思考题

- ❖ 某企业的工会宣称存在性别歧视：女性的收入比男性低，并且这种差异在年龄较大的女性群体表现地尤为明显。
 - 你能设计一个实证检验方案吗？
 - 你打算收集那些数据？

经济学中的实验方法

❖ 随机分配实验

- 处理组(treatment group): 参与实验人员
- 对照组(comparison group)或控制组(control group): 未参与实验的人员

❖ 自然实验(natural experiment)

- 观测值自然产生, 由外生事件引起
- 自然事件、政策变动

第八讲 虚拟变量



思考题

- ❖ 2003年美国有10个州增加了对香烟的税收，预期增税将导致烟草消费量减少。这些州实现了降低烟草消费的目的吗(Table16-2)?
 - 你能设计一个实证检验方案吗?
 - ✓ 收集10个州在增税前后的相关数据，进行纵向对比检验?
 - ✓ 在某一个时点上收集这10个州与非增税州的相关数据，进行横向比较?
 - ✓ 同时收集增税州与非增税州的相关数据，进行比较?

第八讲 虚拟变量



作业

❖ 第七章 (P128) : 习题1、3、5、8