本节内容：

专题：虚拟变量回归

- 5.1 方差分析(ANOVA) 模型
- 5.2 虚拟变量的设置规则
- 5.3 含有两个定性变量的ANOVA 模型
- 5.4 同时含有定性和定量变量的ANOVA 模型
- 5.5 邹检验的虚拟变量方法
- 5.6 分段线性回归：虚拟变量的方法
- 5.7 虚拟变量的引入方式
- 5.8 一个总结性案例

知识回顾

- **变量类型：**
  - 定量变量：可直接测度、数值性因素的度量。（数值型变量）
  - 定性变量：对某种属性存在与否的非数值性度量。（顺序型变量、分类型变量）
- **测量尺度类型：**比率尺度、区间尺度、序数尺度、名义尺度
- **对应关系表：**

测量尺度的类型、属性及变量类型对应表

		测量尺度属性			变量类型
		顺序属性	距离属性	比率属性	
测量尺度类型	比率尺度	√	√	√	数值型变量
	区间尺度	√	√	×	
	序列尺度	√	×	×	顺序型变量
	名义尺度	×	×	×	分类型变量

注：“√”表示测量尺度**具备**该属性，“×”表示测量尺度**不具备**该属性。

方差分析(ANOVA) 模型
——虚拟变量的定义

- **虚拟变量的定义：**
  - 定性变量(qualitative variables)
    - 包括有序类别（顺序变量）和无序类别（分类变量）
    - **用于无法直接定量度量的属性描述：**
      - 顺序型：产品等级、教育程度、满意程度；（顺序取值）
      - 分类型：性别、种族、肤色、宗教、国籍、地区、政治动乱和党派等。（0-1取值）

    提问：定性变量怎样表达出来？如何数量化？
  - 虚拟变量（Dummy Variables）
    - 将取值为0和1的定性变量称为**虚拟变量**。（分类变量）
    - 0-1取值：1—表示具备某种属性，0—表示不具备该属性。
      - 性别(男/女)：D_1=1(男)；D_1=0(女)
      - 肤色(黄/白/黑)：D_1=1(黄)；D_1=0(非黄种人)
        D_2=1(白)；D_2=0（非白种人）

方差分析(ANOVA) 模型
——虚拟变量的性质

- **虚拟变量的性质：**
  - 实质上就是一个将数据区分为相互排斥类别(如男性或女性)的工具。
  - 回归模型中的所有解释变量都是**虚拟变量**，这种模型被称为方差分析(ANOVA) 模型。

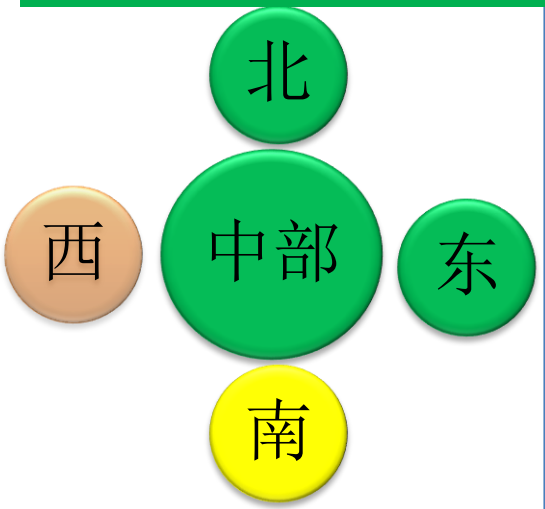
思考1：直接在回归模型中加入定性因素存在哪些困难？
思考2：是否可将这些定性因素进行量化，以达到定性因素能与定量因素有着相同作用之目的。

方差分析(ANOVA) 模型
——说明案例：美国三个区域公立学校教师的平均薪水

表 9—1　2005—2006 年公立学校教师的平均薪水

	薪水	支出	D_2	D_3		薪水	支出	D_2	D_3
康涅狄格	60 822	12 436	1	0	佐治亚	49 905	8 534	0	1
伊利诺伊	58 246	9 275	1	0	肯塔基	43 646	8 300	0	1
印第安纳	47 831	8 935	1	0	路易斯安那	42 816	8 519	0	1
艾奥瓦	43 130	7 807	1	0	马里兰	56 927	9 771	0	1
堪萨斯	43 334	8 373	1	0	密西西比	40 182	7 215	0	1
缅因	41 596	11 285	1	0	北卡罗来纳	46 410	7 675	0	1
马萨诸塞	58 624	12 596	1	0	俄克拉何马	42 379	6 944	0	1
密歇根	54 895	9 880	1	0	南卡罗来纳	44 133	8 377	0	1
明尼苏达	49 634	9 675	1	0	田纳西	43 816	6 979	0	1
密苏里	41 839	7 840	1	0	得克萨斯	44 897	7 547	0	1
内布拉斯加	42 044	7 900	1	0	弗吉尼亚	44 727	9 275	0	1
新罕布什尔	46 527	10 206	1	0	西弗吉尼亚	40 531	9 886	0	1
新泽西	59 920	13 781	1	0	阿拉斯加	54 658	10 171	0	0
纽约	58 537	13 551	1	0	亚利桑那	45 941	5 585	0	0
北达科他	38 822	7 807	1	0	加利福尼亚	63 640	8 486	0	0
俄亥俄	51 937	10 034	1	0	科罗拉多	45 833	8 861	0	0
宾夕法尼亚	54 970	10 711	1	0	夏威夷	51 922	9 879	0	0
罗得岛	55 956	11 089	1	0	爱达荷	42 798	7 042	0	0
南达科他	35 378	7 911	1	0	蒙大拿	41 225	8 361	0	0
佛蒙特	48 370	12 475	1	0	内华达	45 342	6 755	0	0
威斯康星	47 901	9 965	1	0	新墨西哥	42 780	8 622	0	0
阿拉巴马	43 389	7 706	0	1	俄勒冈	50 911	8 649	0	0
阿肯色	44 245	8 402	0	1	犹他	40 566	5 347	0	0
特拉华	54 680	12 036	0	1	华盛顿特区	47 882	7 958	0	0
哥伦比亚特区	59 000	15 508	0	1	怀俄明	50 692	11 596	0	0
佛罗里达	45 308	7 762	0	1					

D_2 =1 位于中/东/北部;
D_2 =0 表示其他地区.



D_3=1 位于美国南部;
D_3=0 表示其他地区.

(东/北/中部)49 538.71 美元
(南部)46 293. 59 美元
(西部)48 104.62 美元

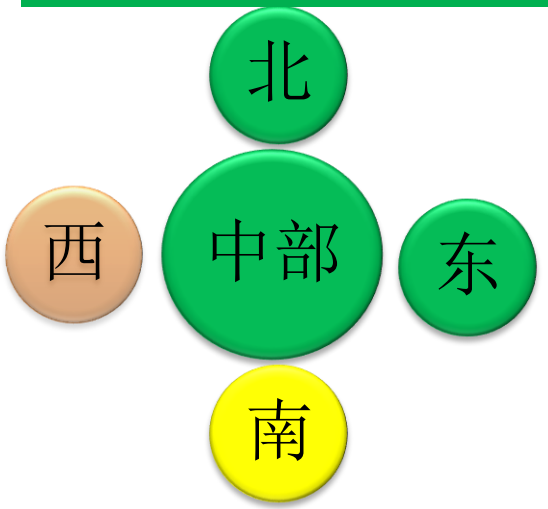
它们在统计上也彼此不同吗?

方差分析(ANOVA) 模型
——说明案例：美国三个区域公立学校教师的平均薪水

- 设总体回归模型PRM：

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i$$

D_2 =1 位于东/中/北部;
D_2 =0 表示其他地区.



D_3=1 位于美国南部;
D_3=0 表示其他地区.

东/北/中部公立学校教师薪水的均值为：

$$E(Y_i \mid D_{2i} = 1, D_{3i} = 0) = \beta_1 + \beta_2$$

南部公立学校教师薪水的均值为：

$$E(Y_i \mid D_{2i} = 0, D_{3i} = 1) = \beta_1 + \beta_3$$

西部教师薪水的均值：

$$E(Y_i \mid D_{2i} = 0, D_{3i} = 0) = \beta_1$$

方差分析(ANOVA) 模型
——说明案例：美国三个区域公立学校教师的平均薪水

$$\hat{Y}_i = 48\,014.615 + 1\,524.099\,D_{2i} - 1\,721.027\,D_{3i}$$

se=(1 857.024) (2 363.139) (2 467.151) R^2=0.044 0

t=(25.853) (0.645) (−0.698)

(0.000 0) (0.522 0) (0.488 8)

D_2 =1 位于中/东/北部;
D_2 =0 表示其他地区.

提问：三个区域的平均薪水具有统计上的显著差异吗？

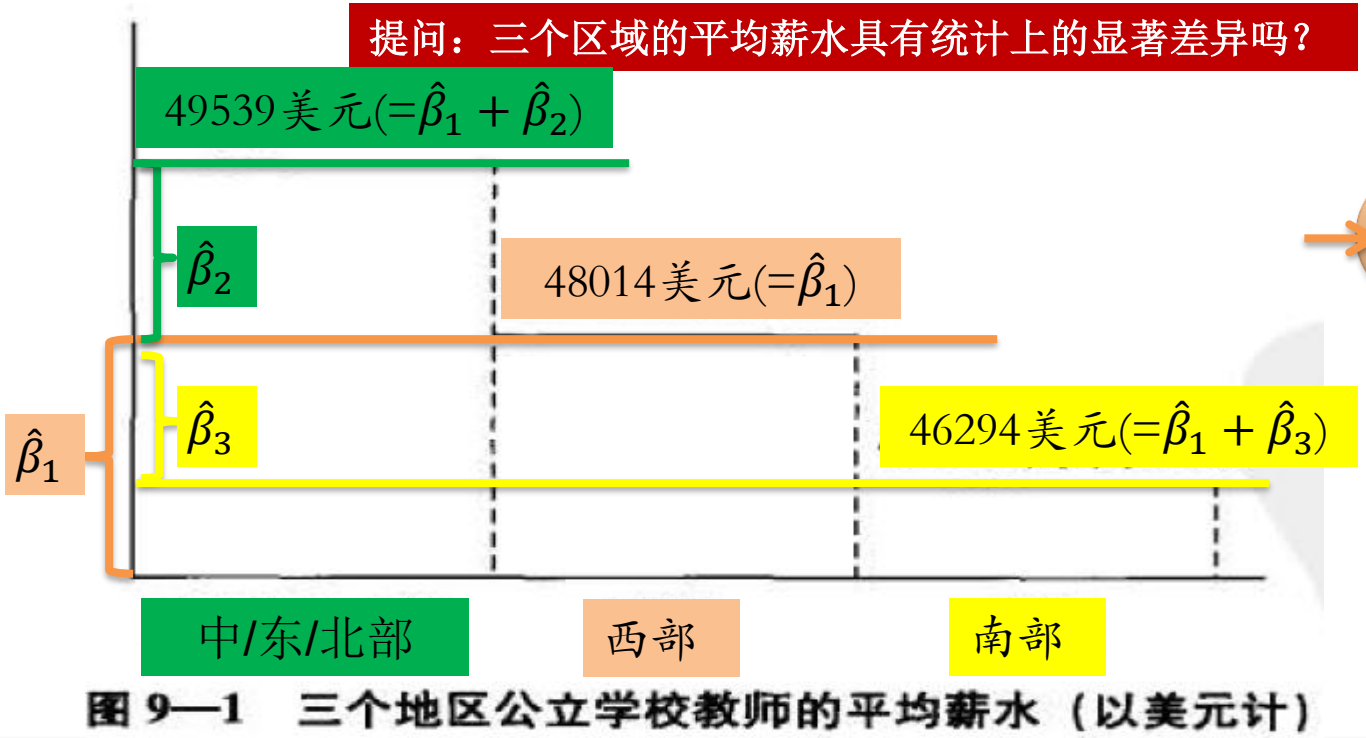
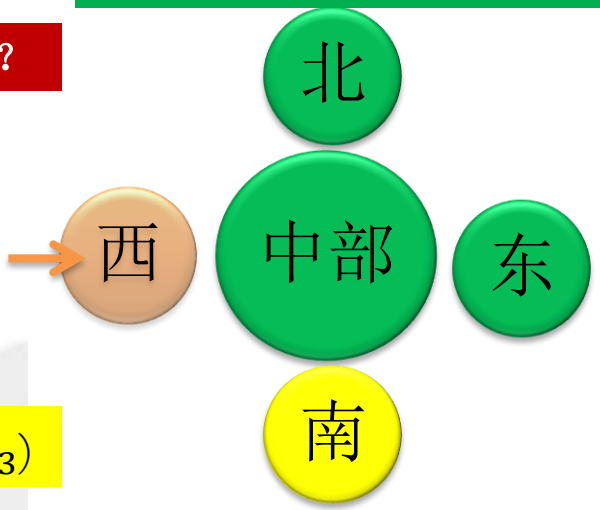


图 9—1 三个地区公立学校教师的平均薪水（以美元计）



D_3=1 位于美国南部;
D_3=0 表示其他地区.

方差分析(ANOVA) 模型
——回归系数的涵义

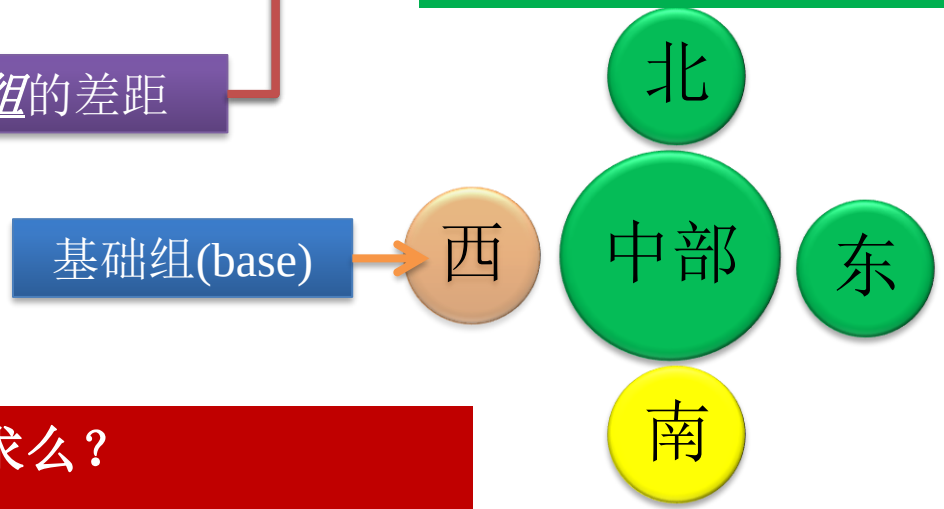
$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i$$

截距系数——代表**基础组**的平均水平

极差截距系数——代表*比较组*与*基础组*的差距

D_2 =1 位于中/东/北部;
D_2 =0 表示其他地区.

基础组(base)



D_3=1 位于美国南部;
D_3=0 表示其他地区.

提问：基础组该怎样确定？有什么要求么？
基础组设定建议：
1.通常设定为平均水平最低或最高的组；
2.根据研究需要，有助于突出目标组与基准组的差异；
3.研究重点关注的组不宜设为基准组。

虚拟变量的设置规则
——模型及数量1

若定性因素具有m个相互排斥属性(或几个水平)：

- **规则1：**
  - 当回归模型有截距项时，只能设(m-1)个虚拟变量；
- **规则2：**
  - 当回归模型无截距项时，则可引入m个虚拟变量。

  特别注意：当模型有截距项时，再引入m个虚拟变量，就会陷入“虚拟变量陷阱”，即所谓的完全共线性。（为什么？）
- **规则3：**
  - 在虚拟变量的设置中：
    - 虚拟变量取值仅为：0或1：
      - 未指定虚拟变量的一组称为基准组，赋值为0；
      - 其他指定虚拟变量的组与基准组相比较，赋值为1；
    - 当存在多个虚拟变量组时：

      正确设定：分次命名多个虚拟变量，取值均为0或1；

      错误设定：命名一个虚拟变量，基准组和其他组取值为：0、1、2

思考：规则1和规则2分别建立虚拟变量回归模型，哪种更好呢？

虚拟变量的设置规则
——“虚拟变量陷阱”

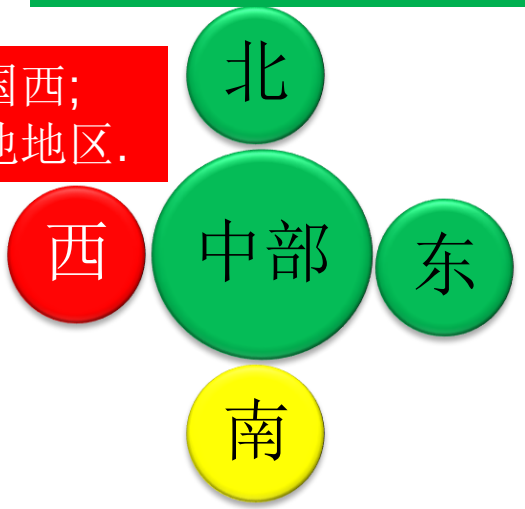
➢ 能否用三个虚拟变量来区分这三个区域呢?

- 建模一：（正确模型）

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i$$

D_2 =1 位于中/东/北部;
D_2 =0 表示其他地区.

D_1=1 位于美国西;
D_1=0 表示其他地区.



- 建模二：（错误模型）

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 D_{1i} + u_i$$

完全共线性(perfect collinearity)！！

- 建模三：（正确模型）

$$Y_i = \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 D_{1i} + u_i$$

D_3=1 位于美国南部;
D_3=0 表示其他地区.

β_2 =东北和中北部教师薪水的均值
β_3 =南部教师薪水的均值
β_4 =西部教师薪水的均值

提问：模型1和模型3的回归系数涵义是一样的么？

虚拟变量的设置规则
——案例说明：何时才采用无截距模型形式？

	薪水	支出	D_2	D_3	D_1
康涅狄格	60 822	12 436	1	0	0
伊利诺伊	58 246	9 275	1	0	0
印第安纳	47 831	8 935	1	0	0
艾奥瓦	43 130	7 807	1	0	0
堪萨斯	43 334	8 373	1	0	0
缅因	41 596	11 285	1	0	0
马萨诸塞	58 624	12 596	1	0	0
密歇根	54 895	9 880	1	0	0
明尼苏达	49 634	9 675	1	0	0
密苏里	41 839	7 840	1	0	0
内布拉斯加	42 044	7 900	1	0	0
新罕布什尔	46 527	10 206	1	0	0
新泽西	59 920	13 781	1	0	0
纽约	58 537	13 551	1	0	0
北达科他	38 822	7 807	1	0	0
俄亥俄	51 937	10 034	1	0	0
宾夕法尼亚	54 970	10 711	1	0	0
罗得岛	55 956	11 089	1	0	0
南达科他	35 378	7 911	1	0	0
佛蒙特	48 370	12 475	1	0	0
威斯康星	47 901	9 965	1	0	0
阿拉巴马	43 389	7 706	0	1	0
阿肯色	44 245	8 402	0	1	0
特拉华	54 680	12 036	0	1	0
哥伦比亚特区	59 000	15 508	0	1	0
佛罗里达	45 308	7 762	0	1	0
佐治亚	49 905	8 534	0	1	0
肯塔基	43 646	8 300	0	1	0
路易斯安那	42 816	8 519	0	1	0
马里兰	56 927	9 771	0	1	0
密西西比	40 182	7 215	0	1	0
北卡罗来纳	46 410	7 675	0	1	0
俄克拉何马	42 379	6 944	0	1	0
南卡罗来纳	44 133	8 377	0	1	0
田纳西	43 816	6 979	0	1	0
得克萨斯	44 897	7 547	0	1	0
弗吉尼亚	44 727	9 275	0	1	0
西弗吉尼亚	40 531	9 886	0	1	0
阿拉斯加	54 658	10 171	0	0	1
亚利桑那	45 941	5 585	0	0	1
加利福尼亚	63 640	8 486	0	0	1
科罗拉多	45 833	8 861	0	0	1
夏威夷	51 922	9 879	0	0	1
爱达荷	42 798	7 042	0	0	1
蒙大拿	41 225	8 361	0	0	1
内华达	45 342	6 755	0	0	1
新墨西哥	42 780	8 622	0	0	1
俄勒冈	50 911	8 649	0	0	1
犹他	40 566	5 347	0	0	1
华盛顿特区	47 882	7 958	0	0	1
怀俄明	50 692	11 596	0	0	1

$$Y_i = 48\,014.62\,D_{1i} + 49\,538.71\,D_{2i} + 46\,293.59\,D_{3i}$$

$$se = (1\,857.204) \quad (1\,461.240) \quad (1\,624.077)$$

$$t = (25.853)^{*} \quad (33.902)^{*} \quad (28.505)^{*} \qquad R^2 = 0.044$$

含有两个定性变量的ANOVA 模型
——说明案例：工资(Y)与居住地和婚姻的关系

Y——小时工资(美元);
D_2——婚姻状况: 1 = 已婚， 0= 其他;
D_3—— 居住地: 1= 南部， 0= 其他.

$$\hat{Y}_i = 8.814\,8 + 1.099\,7\,D_{2i} - 1.672\,9D_{3i}$$

se=	(0.401 5)	(0.464 2)	(0.485 4)
t=	(21.952 8)	(2.368 8)	(−3.446 2)
	(0.000 0)*	(0.018 2)*	(0.000 6)*

R^2 =0.032 2

提问1：大白话解释上述回归函数的结论！
思考1：基准组是什么呢?
思考2：与基组相比在统计上有差异吗?

未婚的非南部居民组

同时含有定性和定量回归元的ANCOVA模型

——协方差分析(analysis of covariance) 模型

- 协方差分析(analysis of covariance) 模型：

是对ANOVA 模型的推广，在一个同时包括定量和定性回归元的模型中，这种模型提供了一种方法，能在统计上控制定量回归元——又被称为协变量(covariates)或控制变量(control variables)——的影响。

同时含有定性和定量回归元的ANCOVA模型

——案例说明：教师薪水(Y)与地区(D)及生均拨款(X)的关系

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 X_i + u_i$$

Y_i=公立学校教师的州平均薪水(美元)
D_{2i} =1，若该州位于东/中/北部;
D_{2i} =0 ，其他
D_{3i} =1，若该州位于南部;
D_{3i} =0，其他。
X_i =对公立学校每个学生的支出(美元);

$$\hat{Y}_i = 28\,694.918 - 2\,954.127 D_{2i} - 3\,112.194 D_{3i} + 2.340\,4 X_i$$

se=(3 262.521) (1 862.576) (1 819.873) (0.359 2)

t=(8.795) (−1.586) (−1.710) (6.515)

R^2=0.497 7

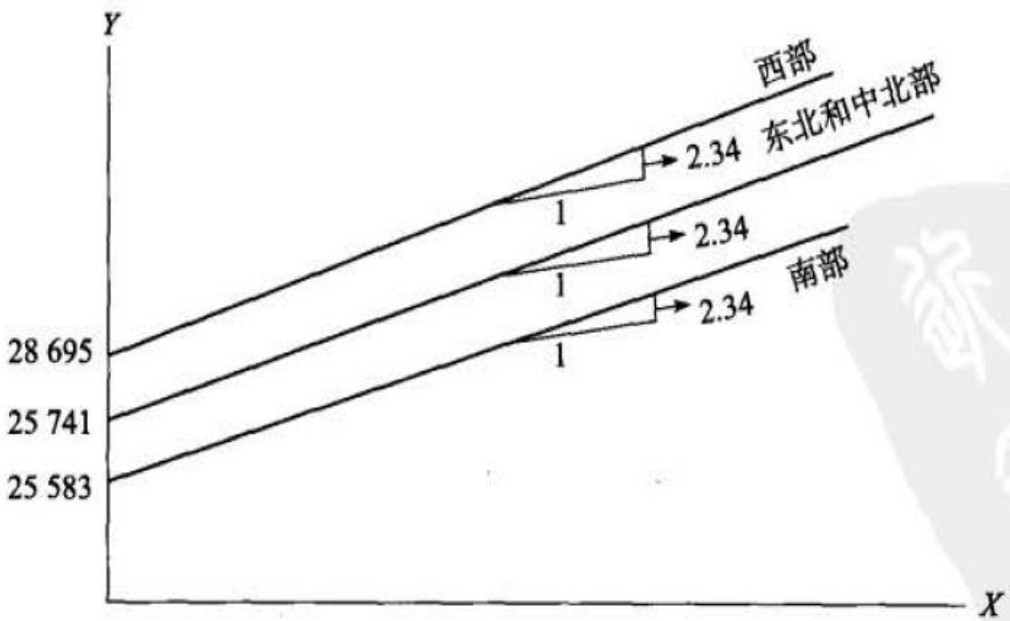


图 9—2 公立学校教师薪水（Y）与对每个学生教育支出（X）的关系

提问1：大白话解释上述回归函数！
思考1：基准组是什么?谁是协变量？
思考2：三条线为什么是平行的?
思考3：统计上来看，南部线和西部线是不一样的么？

邹至庄检验的虚拟变量方法

检验模型的结构稳定性：邹志庄检验(Chow test)

- 原理：检验**时间序列模型**在某个**时间点**前后是否存在结构变动（截距和斜率的变动）。（F检验）

例子：美国1970-1995年间储蓄和收入的关系

邹检验证明：1982年是结构断裂点(Structural Breakpiont)

因此，储蓄和收入之间的关系可以划分为：

$$\text{时期 } 1970\text{—}1981: Y_t = \lambda_1 + \lambda_2 X_t + u_{1t} \quad n_1 = 12$$

$$\text{时期 } 1982\text{—}1995: Y_t = \gamma_1 + \gamma_2 X_t + u_{2t} \quad n_2 = 14$$

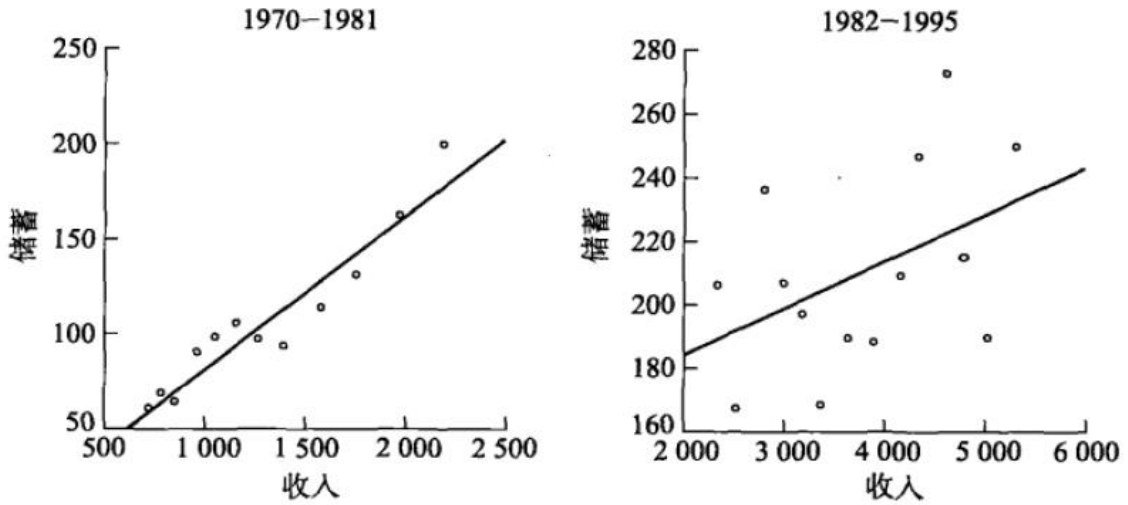


图 8—3

邹至庄检验的虚拟变量方法

邹检验的再讨论1：

例子：美国1970-1995年间储蓄和收入的关系

四种可能的形式：

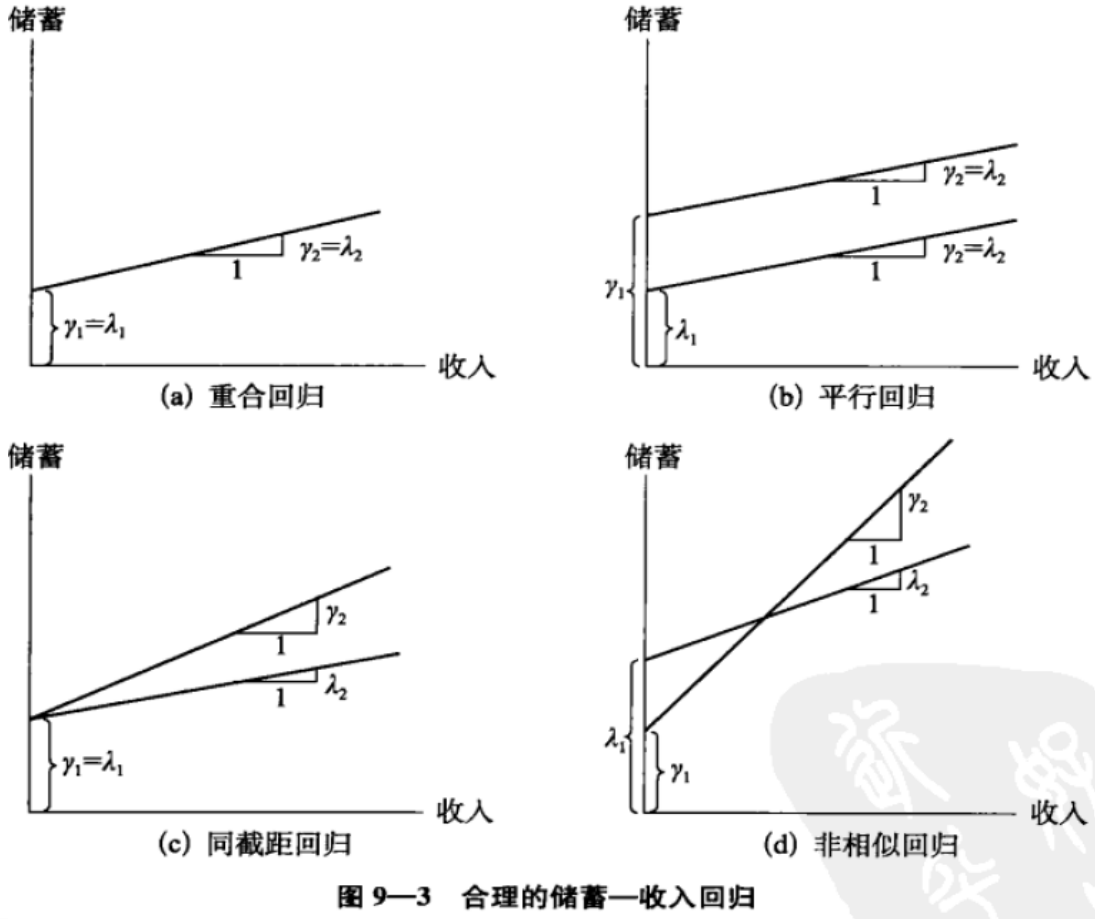


图 9—3 合理的储蓄—收入回归

邹至庄检验的虚拟变量方法

邹检验的再讨论2：

例子：美国1970-1995年间储蓄和收入的关系

虚拟变量的模型表达形式：

$$Y_t = \alpha_1 + \alpha_2 D_t + \beta_1 X_t + \beta_2 (D_t X_t) + u_t$$

其中Y=储蓄；

X=收入；

t=时间；

D =1，1982—1995 年之间的观测；

=0，其他（即 1970—1981 年之间的观测）。

原理：通过检验系数α_2和β_2的显著性（t检验或Wald检验），来实现和邹检验相同的效果。

虚拟变量方法与邹检验的比较：

- 虚拟变量——单一方程；邹检验——两个方程；
- 虚拟变量——t和wald检验；邹检验——F检验；
- 虚拟变量——可以获得截距和斜率的精确形式；
  邹检验——只能判定是否存在截距和斜率的差异
- 虚拟变量估计的自由度和精度高于邹检验。

分段线性回归：虚拟变量方法

分段线性回归的虚拟变量方法

原理：通过设定虚拟变量的方式，检验截面数据再某个临界点前后是否存在结构变动（截距和斜率的变动）。

例子：销售佣金和销售量之间的关系

$$Y_i = \alpha_1 + \beta_1 X_i + \beta_2 (X_i - X^*) D_i + u_i$$

其中Y_i=销售佣金；

X_i=销售员带来的销售量；

X^*=销售临界值，也被称为**结点**（knot）（事先已知）①；

D_i=1，若 $X_i > X^*$；

=0，若 $X_i < X^*$。

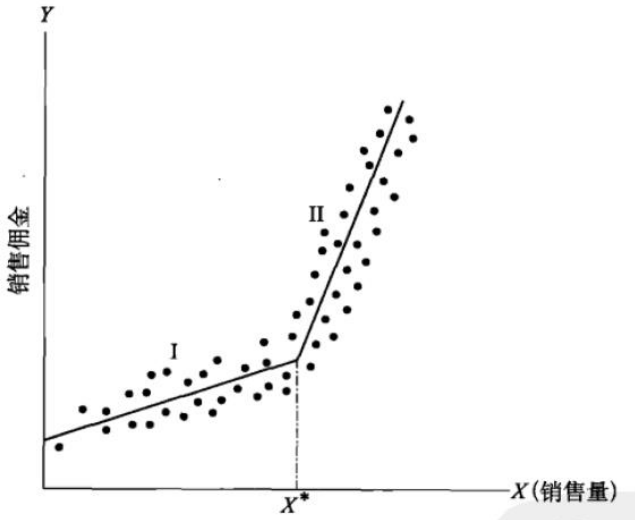


图 9—5　假想销售佣金与销售量之间的关系

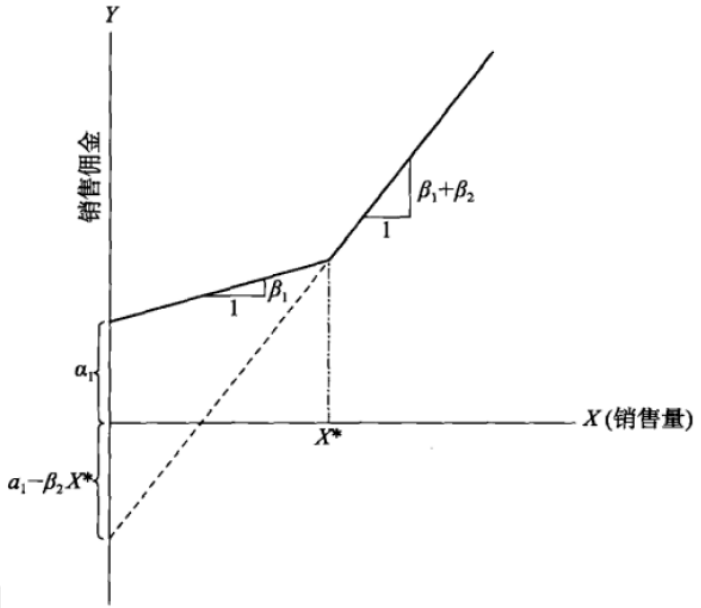


图 9—6　分段线性回归的参数

虚拟变量的引入方式
——加法模型

Y= 小时工资(美元);
X=受教育水平(读书年数)
D_2 = 1，女性； D2 = 0 ，其他
D_3 = 1，黄种人; D3 = 0 ，其他。

若男性工资的均值比女性高，则不论是对哪一个种族来讲都是如此。
若非黄种人的工资均值较高，则不论他们是男性还是女性都是如此。

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta_4 X_i + u_i$$

$$E(Y_i \mid D_{2i} = 0, D_{3i} = 0, X_i) = \beta_1 + \beta_4 X_i$$ 男性非黄种人的时均工资

$$E(Y_i \mid D_{2i} = 1, D_{3i} = 1, X_i) = (\beta_1 + \beta_2 + \beta_3) + \beta_4 X_i$$ 女性黄种人的时均工资

$$E(Y_i \mid D_{2i} = 1, D_{3i} = 0, X_i) = (\beta_1 + \beta_2) + \beta_4 X_i$$ 女性非黄种人的时均工资

$$\hat{Y}_i = -0.261\,0 - 2.360\,6D_{2i} - 1.732\,7D_{3i} + 0.802\,8X_i$$
$$t = (-0.235\,7) \quad (-5.487\,3) \quad (-2.180\,3) \quad (9.909\,4)$$
$$R^2 = 0.203\,2, \quad n = 528$$

提问1：如何解释回归结果？

虚拟变量的引入方式
——乘法模型

Y= 小时工资(美元);
X=受教育水平(读书年数)
D_2 = 1， 女性； D2 = 0 ， 其他
D_3 = 1， 黄种人; D3 = 0 ， 其他。

若女性工资低，但女性黄种人可能比女性非黄种人工资高。
若黄种人工资低，但黄种人女性可能比黄种人男性工资高。

$$Y_i = \beta_1 + \beta_2 D_{2i} + \beta_3 D_{3i} + \beta^* (D_{2i} D_{3i}) + \beta_4 X_i + u_i$$

$$E(Y_i \mid D_{2i} = 0, D_{3i} = 0, X_i) = \beta_1 + \beta_4 X_i$$

男性非黄种人的时均工资

$$E(Y_i \mid D_{2i} = 1, D_{3i} = 1, X_i) = (\beta_1 + \beta_2 + \beta_3 + \beta^*) + \beta_4 X_i$$

女性黄种人的时均工资

$$\hat{Y}_i = -0.261\,0 - 2.360\,6 D_{2i} - 1.732\,7 D_{3i} + 2.128\,9 D_{2i} D_{3i} + 0.802\,8 X_i$$
$$t = (-0.235\,7)^{**} \quad (-5.487\,3)^{*} \quad (-2.180\,3)^{*} \quad (1.742\,0)^{**} \quad (9.909\,5)^{**} \quad (9.6.5)$$
$$R^2 = 0.203\,2, \quad n = 528$$

提问1：如何解释回归结果？

一个总结性案例
——来自印度261个工人的样本数据

表 9—7　　1990 年的一个印度工人样本

WI	AGE	DE_2	DE_3	DE_4	DPT	D_{SEX}	WI	AGE	DE_2	DE_3	DE_4	DPT	D_{SEX}
120	57	0	0	0	0	0	120	21	0	0	0	0	0
224	48	0	0	1	1	0	25	18	0	0	0	0	1
132	38	0	0	0	0	0	25	11	0	0	0	0	1
75	27	0	1	0	0	0	30	38	0	0	0	1	1
111	23	0	1	0	0	1	30	17	0	0	0	1	1
127	22	0	1	0	0	0	122	20	0	0	0	0	0
30	18	0	0	0	0	0	288	50	0	1	0	1	0
24	12	0	0	0	0	0	75	45	0	0	0	0	1
119	38	0	0	0	1	0	79	60	0	0	0	0	0
75	55	0	0	0	0	0	85. 3	26	1	0	0	0	1
97	25	0	1	0	0	0	325	55	0	0	0	1	0
150	36	0	0	0	0	0	121	27	0	1	0	0	0
25	28	0	0	0	0	1	600	35	1	0	0	0	0
15	13	0	0	0	0	1	52	19	0	0	0	0	0
131	55	0	0	0	0	0	117	28	1	0	0	0	0

WI= 周收入(卢比)

Age=年龄

DE_2= 1,受过初等教育
DE_2= 0,未受过初等教育
DE_3 =1,受过中等教育
DE_3= 0,未受过中等教育
DE_4 =1,受过高等教育
DE_4= 0,未受过高等教育

DPT = 1, 永久性工作
DPT =0 , 暂时性工作

DSEX = 1,男性工人
DSEX = 0,女性工人

$$\ln WI_i = \beta_1 + \beta_2 AGE_i + \beta_3 D_{SEX} + \beta_4 DE_2 + \beta_5 DE_3 + \beta_6 DE_4 + \beta_7 DPT + u_i$$

一个总结性案例
——案例说明：加法模型

Dependent Variable: Ln(WI)
Method: Least Squares
Sample: 1 261
Included observations: 261

	Coefficient	Std. Error	t-Statistic	Prob.
C	3.706872	0.113845	32.56055	0.0000
AGE	0.026549	0.003117	8.516848	0.0000
D_{SEX}	-0.656338	0.088796	-7.391529	0.0000
DE_2	0.113862	0.098542	1.155473	0.2490
DE_3	0.412589	0.096383	4.280732	0.0000
DE_4	0.554129	0.155224	3.569862	0.0004
DPT	0.558348	0.079990	6.980248	0.0000

R-squared	0.534969	Mean dependent var.	4.793390
Adjusted R-squared	0.523984	S.D. dependent var.	0.834277
S.E. of regression	0.575600	Akaike info criterion	1.759648
Sum squared resid.	84.15421	Schwarz criterion	1.855248
Log likelihood	-222.6340	Hannan-Quinn criter.	1.798076
F-statistic	48.70008	Durbin-Watson stat.	1.853361
Prob(F-statistic)	0.000000		

一个总结性案例
——乘法模型（性别与受教育程度的交互）

Dependent Variable: Ln(WI)
Method: Least Squares
Sample: 1 261
Included observations: 261

1.性别与受教育程度之间有可能存在交互影响吗?
2.受过高等教育的男性工人比受过高等教育的女性工人挣的周收入更高吗?

	Coefficient	Std. Error	t-Statistic	Prob.
C	3.717540	0.114536	32.45734	0.0000
AGE	0.027051	0.003133	8.634553	0.0000
D_{SEX}	-0.758975	0.110410	-6.874148	0.0000
DE_2	0.088923	0.106827	0.832402	0.4060
DE_3	0.350574	0.104309	3.360913	0.0009
DE_4	0.438673	0.186996	2.345898	0.0198
$D_{SEX}*DE_2$	0.114908	0.275039	0.417788	0.6765
$D_{SEX}*DE_3$	0.391052	0.259261	1.508337	0.1327
$D_{SEX}*DE_4$	0.369520	0.313503	1.178681	0.2396
DPT	0.551658	0.080076	6.889198	0.0000

R-squared	0.540810	Mean dependent var.	4.793390
Adjusted R-squared	0.524345	S.D. dependent var.	0.834277
S.E. of regression	0.575382	Akaike info criterion	1.769997
Sum squared resid.	83.09731	Schwarz criterion	1.906569
Log likelihood	-220.9847	Hannan-Quinn criter.	1.824895
F-statistic	32.84603	Durbin-Watson stat.	1.856488
Prob (F-statistic)	0.000000		

一个总结性案例
——乘法模型(去掉受教育程度变量)

Dependent Variable: LOG(WI)
Method: Least Squares
Sample: 1 261
Included observations: 261

	Coefficient	Std. Error	t-Statistic	Prob.
C	3.836483	0.106785	35.92725	0.0000
AGE	0.025990	0.003170	8.197991	0.0000
D_{SEX}	-0.868617	0.106429	-8.161508	0.0000
$D_{SEX}*DE_2$	0.200823	0.259511	0.773851	0.4397
$D_{SEX}*DE_3$	0.716722	0.245021	2.925140	0.0038
$D_{SEX}*DE_4$	0.752652	0.265975	2.829789	0.0050
DPT	0.627272	0.078869	7.953332	0.0000

R-squared	0.514449	Mean dependent var.	4.793390
Adjusted R-squared	0.502979	S.D. dependent var.	0.834277
S.E. of regression	0.588163	Akaike info criterion	1.802828
Sum squared resid.	87.86766	Schwarz criterion	1.898429
Log likelihood	-228.2691	Hannan-Quinn criter.	1.841257
F-statistic	44.85284	Durbin-Watson stat.	1.873421
Prob (F-statistic)	0.000000		

虚拟变量模型经济含义的解释

- **半对数（线性到对数）模型中系数β的经济含义：**
  - **数值型变量**：解释变量X（在平均意义上）的1单位变动，导致被解释变量Y（在平均意义上）变动β比例（或百分比）的变动。
  - **分类型变量**（虚拟变量）：解释变量X(在平均意义上)1单位变动（从0—>1），导致被解释变量Y的（**中位数**）变动$e^{\beta}-1$单位。(课本P302,P312)
    - 数值模拟证明：反对数的经济含义——从均值到中位数。（课本P297，例子9.8）
    - 虚拟变量的“边际变动”的说明

虚拟变量模型经济含义的解释

- 虚拟变量的“边际变动”的说明（课本P312页）

我们在 9.10 节注意到，在如下形式的模型中

$$\ln Y_i = \beta_1 + \beta_2 D_i \tag{1}$$

对于取值 1 或 0 的虚拟变量，Y 的相对变化（即半弹性）可如下得到：（β_2 估计值的反对数－1）×100，即

$$(e^{\beta_2} - 1) \times 100 \tag{2}$$

证明如下：由于对数和指数互为反函数，所以我们可以把方程（1）写成

$$\ln Y_i = \beta_1 + \ln(e^{\beta_2 D_i}) \tag{3}$$

现在，当 $D=0$ 时，$e^{\beta_2 D_i} = 1$，当 $D=1$ 时，$e^{\beta_2 D_i} = e^{\beta_2}$。因此，从状态 0 到状态 1，$\ln Y_i$ 变化了 $(e^{\beta_2} - 1)$。但一个变量对数的变化只是相对变化，乘以 100 后就得到百分数变化。因此百分数变化就如所要证明的那样为 $(e^{\beta_2} - 1) \times 100$。（注：$\ln_e e = 1$，即以 e 为底 e 的对数等于 1，就像以 10 为底 10 的对数等于 1 一样。记住，以 e 为底的对数被称为自然对数，而以 10 为底的对数被称为常用对数。）