

第04章 扩展方法

4.1 模型函数形式

4.2 虚拟变量模型

4.2 虚拟变量回归模型

4.2.0 相关知识回顾

4.2.1 虚拟变量的设置规则

4.2.2 方差分析(ANOVA) 模型

4.2.3 只含有一个定性变量的ANOVA
模型

4.2.4 同时含有一个定性和一个定量变
量的ANOVA 模型

4.2.5 同时含有多个定性和定量变量的
ANOVA 模型

4.2.6 印度工人工资案例

4.2.7 时间序列季节虚拟变量模型

4.2.8 分段线性回归模型

4.2.0 相关知识回顾



变量类型

定量变量（Quantitative variable）一般也称为连续变量，是由测量或计数、统计所得到的量，可以通过数值表达，并具有直接的数值含义。

定性变量（qualitative variables）：又被称为指标变量（indicator variables）、分类变量（categorical variables），主要用于区分事物性质差异，往往用语义类别表达，没有直接的数值含义。

- 性别（男；女）
- 肤色（黄色；白色；黑色；其他）
- 种族、宗教、国籍、地区、政治动乱和党派等。

| 提问：定性变量怎样表达出来？如何数量化？



变量尺度

变量尺度（Variable scale）：刻画的是变量的数值含义或数值关系。它将意味着在数值含义和关系上，变量是有层次级别的差异性。根据变量层级不同，具体可以分为由低到高的4个层级：

- **名义尺度（nominal scale）** 变量：这类变量只用于属性分类，不具备任何数值含义或数值关系，也即不能加、减、乘、除，也不能比较大小。
- **序数尺度（order scale）** 变量：这类变量具备很少的数值含义或数值关系，它可以比较大小，但不能进行加、减、乘、除。
- **区间尺度（interval scale）** 变量：这类变量具备一定的数值含义或数值关系，它可以比较大小，也可以进行加、减，但不能进行乘、除。
- **比率尺度（ratio scale）** 变量：这类变量具备最多的数值含义或数值关系，它可以比较大小，也可以进行加、减、乘、除。



区域经理年薪案例(数据)

变量类型和变量尺度

salary	sale	score	race
25	140	2	yellow
35	195	3	white
27	184	1	yellow
42	256	5	white
38	207	4	black

区域经理年薪案例中，公司有五名区域经理，分别负责不同的国际市场。

定量变量 *salary* 表示区域经理的年薪（万元）；变量 *sale* 表示负责市场的销售额；变量 *score* 表示客户对区域经理的评价（1表示很不满意，2表示不满意，3表示一般，4表示很满意，5表示非常满意）；变量 *race* 表示区域市场主要消费群体的肤色（*yellow* 表示黄色消费群体、*white* 表示白色消费群体，*black* 表示黑色消费群体）。



区域经理年薪案例(变量)

变量类型和变量尺度

salary	sale	score	race
25	140	2	yellow
35	195	3	white
27	184	1	yellow
42	256	5	white
38	207	4	black

根据以上定义，区域经理年薪案例中，可认为年薪 *salary*、销售额 *sale*，以及客户评价 *score*为定量变量，消费群体主要肤色 *race*为定性变量。从变量的度量尺度来看：

- 年薪 *salary*和销售额 *sale*两个变量为比率尺度变量
- 客户评价 *score*变量为序数尺度变量
- 消费群体主要肤色 *race*为名义尺度变量

4.2.1 虚拟变量的设置规则



定性变量对回归模型的影响

计量经济学建模分析中，我们常常需要把一些定性变量（Qualitative variables）（如性别、地区、党派等）作为自变量放入回归模型中。从变量层次（Variable Scale）来看，这些变量没有具体的取值，只有特定属性类别。

显然，诸如此类的变量如果直接放到线性回归模型中，将会产生一系列的参数估计、模型解释等问题。

$$\text{salary}_i = \beta_1 + \beta_2 \text{sale}_i + \beta_3 \text{score}_i + \beta_4 \text{race}_i + u_i$$

- 一个定性变量的不同数据取值，称为该定性变量的属性。
- 定性变量的任一属性，都可以设置为一个虚拟变量。
- 我们可以用一套虚拟变量体系来完全表达一个定性变量。
- 按照一定规则构建虚拟变量回归模型，避免参数估计、模型解释等问题的出现。



虚拟变量的定义

虚拟变量 (dummy variable) : 将取值为0和1的人造变量称为虚拟变量。

- 对定性变量的量化可采用虚拟变量的方式实现。
- 一般而言, 1表示出现 (或具备) 某种属性, 0表示没有 (或不具备) 某种属性。

对于某定性变量的任一特定属性, 可以构造出一个虚拟变量 (记为D), 使得该虚拟变量能够表达这一属性。同时, 给该虚拟变量D赋值为1, 记为具备这一属性; 给该虚拟变量赋值为0, 记为不具备该属性。

正式地, 假设定性变量 X 具有 m 个属性 a_1, a_2, \dots, a_m , 对于任意属性 k , ($k \in 1, 2, \dots, m$), 可以定义如下的虚拟变量 D_k :

$$D_k = \begin{cases} 1, & \text{if } a_k \\ 0, & \text{if not } a_k \end{cases}$$



区域经理年薪案例（虚拟变量）

区域经理年薪案例中，定性变量 `race`（人种，其取值为黄种人/白种人/黑种人），可以构造出3个虚拟变量

$$race \{a_1 = \text{yellow}, a_2 = \text{white}, a_3 = \text{black}\}$$

$$\text{dummy} \Rightarrow \begin{cases} D_1 = \begin{cases} 1, & \text{yellow} \\ 0, & \text{not yellow} \end{cases} \\ D_2 = \begin{cases} 1, & \text{white} \\ 0, & \text{not white} \end{cases} \\ D_3 = \begin{cases} 1, & \text{black} \\ 0, & \text{not black} \end{cases} \end{cases}$$



虚拟变量体系

虚拟变量体系：完整表达某个定性变量全部信息的一组虚拟变量。

正式地，假设定性变量 X 具有 m 个属性 a_1, a_2, \dots, a_m ，可以用如下一组虚拟变量 $D_1, \dots, D_k, \dots, D_m$ 完全表达该定性变量：

$$X\{a_1, a_2, \dots, a_m\} \Rightarrow \begin{cases} D_1 = \begin{cases} 1, & \text{if } a_1 \\ 0, & \text{if not } a_1 \end{cases} \\ \vdots \\ D_k = \begin{cases} 1, & \text{if } a_k \\ 0, & \text{if not } a_k \end{cases} \\ \vdots \\ D_m = \begin{cases} 1, & \text{if } a_m \\ 0, & \text{if not } a_m \end{cases} \end{cases}$$



区域经理年薪案例（虚拟变量体系）

实际数据操作中，一般需要对定性变量 *race* 进行重新编码（recode），生成三个对应的虚拟变量。

把定性变量转变为虚拟变量体系

salary	sale	score	race	race_black	race_white	race_yellow
25	140	2	yellow	0	0	1
35	195	3	white	0	1	0
27	184	1	yellow	0	0	1
42	256	5	white	0	1	0
38	207	4	black	1	0	0

4.2.2 方差分析模型模型 (ANOVA model)



定义

区域经理薪水案例中，如果不区分变量类型和特征，做如下的回归模型，则回归分析结果将会带来严重的问题。

$$salary_i = \beta_1 + \beta_2 sale_i + \beta_3 score_i + \beta_4 race_i + u_i$$

事实上，应该将上述模型转换为虚拟变量回归模型（Dummy model）。

$$salary_i = \beta_1 + \beta_2 sale_i + \beta_3 score_i + \beta_4 race_yellow_i + \beta_5 race_white_i + u_i$$

$$salary_i = \beta_2 sale_i + \beta_3 score_i + \beta_4 race_yellow_i + \beta_5 race_white_i + \beta_6 race_black_i + u_i$$



定义

一个线性回归模型，只要回归元中包含了虚拟变量，这种模型就被称为虚拟变量回归模型，也可以称为方差分析模型（Analysis of variance, ANOVA）。

方差分析模型（Analysis of variance, ANOVA）常用来分析定量化的因变量 Y 与定性回归元或虚拟变量之间的统计显著性关系。一般是通过比较不同类别或不同组的均值差，例如采用t检验可以判断两组均值是否有显著的差异。

提问：你还能不能设置成其他类型的模型形式？怎样设置才是正确的方差分析模型？



方差分析模型：本质

$$salary_i = \beta_1 + \beta_2 sale_i + \beta_3 score_i + \beta_4 race_yellow_i + \beta_5 race_white_i + u_i$$

很显然，在上述总体回归模型下，可以得到所有3类“分组”情形下的期望年薪水水平：

$$\begin{aligned} E(Y | race_yellow = 1, race_white = 0, sale, score) \\ = \beta_1 + \beta_2 sale + \beta_3 score + \beta_4 \end{aligned} \quad (\text{market yellow})$$

$$\begin{aligned} E(Y | race_yellow = 0, race_white = 1, sale, score) \\ = \beta_1 + \beta_2 sale + \beta_3 score + \beta_5 \end{aligned} \quad (\text{market white})$$

$$\begin{aligned} E(Y | race_yellow = 0, race_white = 0, sale, score) \\ = \beta_1 + \beta_2 sale + \beta_3 score \end{aligned} \quad (\text{market black})$$



方差分析模型：内涵

$$salary_i = \beta_1 + \beta_2 sale_i + \beta_3 score_i + \beta_4 race_yellow_i + \beta_5 race_white_i + u_i$$

上述模型被称其为有截距的含有虚拟变量的、加法形式的回归模型。显然，虚拟变量 $race_black$ 没有进入模型中；模型设置有截距项 β_1 。在这种设定下，我们称：

- 黑色(black)为模型的基础组
- 黄色(yellow)和白色(white)分别为模型的比较组。
- 有序变量 $score$ 为协变量(covariates)或控制变量(control variable)
- β_1 为截距系数，代表基础组的期望水平
- β_2, β_3 为平行斜率系数，代表协变量的影响效应
- β_4, β_5 为极差系数，代表的是比较组与基础组期望水平的差距



方差分析模型的类型：数量关系

根据回归元包含定量变量和虚拟变量的数量关系，可以将虚拟变量回归模型分为：

- 只含有虚拟变量的回归模型：全部解释变量都是由虚拟变量构成
- 同时含有虚拟变量和定量变量的回归模型：解释变量同时含有虚拟变量和定量变量



方差分析模型的类型：引入方式

根据模型中虚拟变量引入方式的不同，可以划分为：

- 加法模型：虚拟变量以独立项的形式出现在方程中
- 乘法模型：虚拟变量以交叉项的形式出现在方程中
- 混合模型：虚拟变量以独立项和/或交叉项的形式出现在方程中^[有时候模型设置中，某个虚拟变量体系（用来表达某个定性变量）的独立项可以完全不出现在方程中（也即没有它们的加法形式），而却可以出现它们与其他变量的交叉项（也即可以出现它们与其他变量的乘法形式）。]
 - 完全混合模型
 - 部分混合模型



方差分析模型的类型：基础组

根据虚拟变量模型是否参照基础组，可以划分为：

- 有截距模型：此时模型解释中将有明确的基础组，其他组可以直接与之参照对比。
- 无截距模型：此时模型解释中将没有明确的基础组，各组间将不直接参照对比。



方差分析模型的类型：函数形式

根据模型中的因变量 Y 是否取对数，可以划分为（半对数或对数模型将蕴含着弹性和斜率的经济学含义，在解释虚拟变量回归模型中往往很有现实意义）：

- 经典线性模型：因变量为 Y
- 半对数模型：因变量为 $\ln(Y)$



方差分析模型的类型：应用情景

根据虚拟变量模型应用情景的不同，可以划分为：

- 截面数据虚拟变量回归模型：此时虚拟变量用于表达回归元为定性变量的情形
- 时间序列季节虚拟变量回归模型：此时虚拟变量用于表达季节周期
- 分段线性虚拟变量回归模型：此时虚拟变量用于表达阈值分段



方差分析模型的类型：综合

对于具体的实证分析案例，我们往往需要根据变量的属性和特征，构建不同类型的虚拟变量回归模型，比较不同模型的回归分析结果，甄选并得到其中相对理想的模型。

例如，仅是考虑基础组的有截距模型，可能用到的各类备选组合模型至少包括：

- 只含有虚拟变量的、加法形式的经典回归模型
- 只含有虚拟变量的、加法形式的半对数回归模型
- 只含有虚拟变量的、乘法形式的经典回归模型
- 只含有虚拟变量的、乘法形式的半对数回归模型
- ...
- 同时含有虚拟变量和定量变量的、加法形式的经典回归模型
- 同时含有虚拟变量和定量变量的、加法形式的半对数回归模型
- 同时含有虚拟变量和定量变量的、乘法形式的经典回归模型
- 同时含有虚拟变量和定量变量的、乘法形式的半对数回归模型

4.2.3 只含有一个定性变量的ANOVA模型



公立学校教师薪水案例

下面我们将以公立学校教师薪水案例，进行分析和论述。

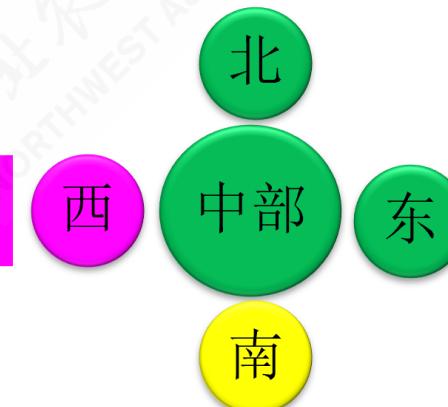


变量

一项研究关注于对美国51个州公立学校教师薪水的分析：

- 变量 $Salary$ 表示公立学校教师的平均薪水； $Spend$ 表示公立学校教师的平均支出；
- $state$ 表示公立学校所在州名称； $Region$ 表示州所属的区位（ $West$ 表示西部州； $M.E.N$ 表示中东北部州； $South$ 表示南部州）。

$D2 = 1$ 位于中/东/北部;
 $D2 = 0$ 表示其他地区.



$D1 = 1$ 位于美国西部;
 $D1 = 0$ 表示其他地区.

$D3 = 1$ 位于美国南部;
 $D3 = 0$ 表示其他地区.



数据

美国三个区域公立学校教师的薪水n=(51)

state	Region	Salary	Spend
Connecticut	M.E.N	60822	12436
Illinois	M.E.N	58246	9275
Indiana	M.E.N	47831	8935
Iowa	M.E.N	43130	7807
Kansas	M.E.N	43334	8373
Maine	M.E.N	41596	11285
Massachusetts	M.E.N	58624	12596
Michigan	M.E.N	54895	9880
Minnesota	M.E.N	49634	9675

Showing 1 to 9 of 51 entries

Previous



描述统计

三个区域公立学校教师的平均薪水n=(51)

N.state	Mean.Salary	Max.Salary	Min.Salary	SD.Salary
21	49,538.71	60,822.00	35,378.00	7,645.47
17	46,293.59	59,000.00	40,182.00	5,543.65
13	48,014.62	63,640.00	40,566.00	6,400.05

根据以上简单的汇总计算：

- 教师的平均薪水：中东北部为49 538.71 美元；南部为46 293.59 美元；西部为48 104.62美元。
- 那么，三个地区的平均薪水在统计上也彼此不同吗？



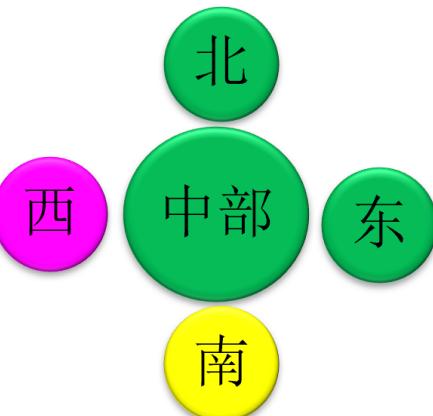
虚拟变量体系

根据前述定义，我们可以将定性变量 *Region* 设置为如下的虚拟变量体系：

Region{*West*; *M.E.N*; *South*}

$$\Rightarrow \begin{cases} D_1 = \begin{cases} 1, & \text{if } West \\ 0, & \text{if not } West \end{cases} \\ D_2 = \begin{cases} 1, & \text{if } M.E.N \\ 0, & \text{if not } M.E.N \end{cases} \\ D_3 = \begin{cases} 1, & \text{if } South \\ 0, & \text{if not } South \end{cases} \end{cases}$$

D2=1 位于中/东/北部;
D2=0 表示其他地区.



D1=1 位于美国西部;
D1=0 表示其他地区.

D3=1 位于美国南部;
D3=0 表示其他地区.



虚拟变量变换

实际建模之前，我们需要把定性变量 *Region* 进行数据变换，得到虚拟变量的数据：

把定性变量Region处理成虚拟变量n=(51)

state	Region	Salary	Spend	D1	D2	D3
Connecticut	M.E.N	60822	12436	0	1	0
Illinois	M.E.N	58246	9275	0	1	0
Indiana	M.E.N	47831	8935	0	1	0
Iowa	M.E.N	43130	7807	0	1	0
Kansas	M.E.N	43334	8373	0	1	0
Maine	M.E.N	41596	11285	0	1	0
Massachusetts	M.E.N	58624	12596	0	1	0

Showing 1 to 7 of 51 entries

Previous



有截距虚拟变量模型：PRM

我们可以构建薪水 (*Salary*) 对区域虚拟变量 ($D2; D3$) 的有截距总体回归模型PRM：

$$Salary_i = \beta_1 + \beta_2 D2_i + \beta_3 D3_i + u_i$$

理论上，我们可以得到三个区域教师薪水的期望值：

$$E(Salary|D2 = 1, D3 = 0) = \beta_1 + \beta_2 \quad (\text{M.E.N})$$

$$E(Salary|D2 = 0, D3 = 1) = \beta_1 + \beta_3 \quad (\text{South})$$

$$E(Salary|D2 = 0, D3 = 0) = \beta_1 \quad (\text{West})$$

$D2 = 1$ 位于中/东/北部;
 $D2 = 0$ 表示其他地区.

北

西

中部

东

南

$D1=1$ 位于美国西部;
 $D1=0$ 表示其他地区.

$D3=1$ 位于美国南部;
 $D3=0$ 表示其他地区.

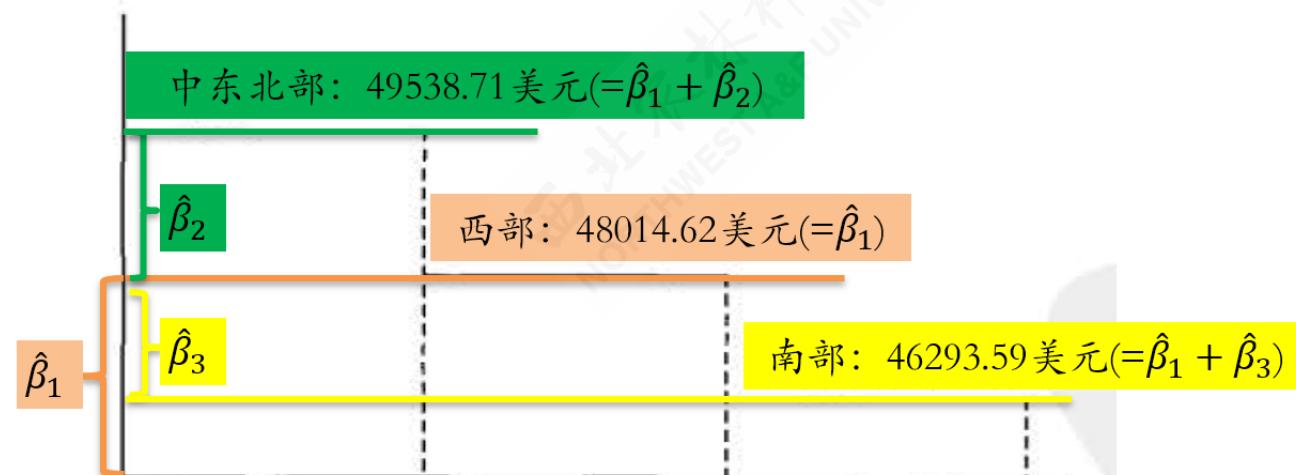


有截距虚拟变量模型：OLS估计

$$\widehat{\text{Salary}} = + 48014.62 + 1524.10D2 - 1721.03D3$$

(t) (25.8532) (0.6449) (-0.6976)
(se) (1857.2037) (2363.1394) (2467.1508)

(fitness) $R^2 = 0.0440; \bar{R}^2 = 0.0041$
 $F^* = 1.10; p = 0.3399$



- 基础组是谁？极差系数的含义？
- 三个区域的平均薪水具有统计上的显著差异吗？



有截距虚拟变量模型：OLS估计

$$\text{Salary}_i = \beta_1 + \beta_2 D2_i + \beta_3 D3_i + u_i$$

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

提问：

- 基础组该怎样确定？
- 有什么要求么？

D2=1 位于中/东/北部;
D2=0 表示其他地区.



D1=1 位于美国西部;
D1=0 表示其他地区.

D3=1 位于美国南部;
D3=0 表示其他地区.



无截距虚拟变量模型：PRM2

我们也可以构建薪水（*Salary*）对区域虚拟变量（ $D1; D2; D3$ ）的无截距总体回归模型PRM：

$$Salary_i = \alpha_1 D1_i + \alpha_2 D2_i + \alpha_3 D3_i + u_i$$

理论上，我们可以得到三个区域教师薪水的期望值：

$$E(Salary|D1 = 0, D2 = 1, D3 = 0) = \alpha_2 \quad (\text{M.E.N})$$

$$E(Salary|D1 = 0, D2 = 0, D3 = 1) = \alpha_3 \quad (\text{South})$$

$$E(Salary|D1 = 1, D2 = 0, D3 = 0) = \alpha_1 \quad (\text{West})$$

D2=1 位于中/东/北部;
D2=0 表示其他地区.

北

西

中部

东

南

D1=1 位于美国西部;
D1=0 表示其他地区.

D3=1 位于美国南部;
D3=0 表示其他地区.



无截距虚拟变量模型：OLS估计

$$\widehat{\text{Salary}} = + 48014.62D1 + 49538.71D2 + 46293.59D3$$
$$(t) \quad (25.8532) \quad (33.9018) \quad (28.5045)$$
$$(\text{se}) \quad (1857.2037) \quad (1461.2400) \quad (1624.0775)$$
$$(\text{fitness}) \quad R^2 = 0.9821; \quad \bar{R}^2 = 0.9810$$
$$F^* = 876.74; \quad p = 0.0000$$



无截距虚拟变量模型：OLS估计

$$\text{Salary}_i = \alpha_1 D1_i + \alpha_2 D2_i + \alpha_3 D3_i + u_i$$

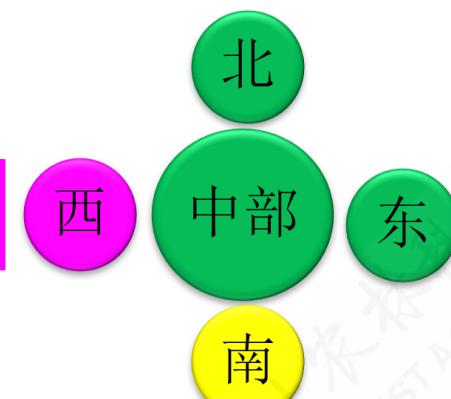


提问：

- 此时谁是基础组？
- 回归模型有没有基础组，很重要么？

D2=1 位于中/东/北部;
D2=0 表示其他地区.

D1=1 位于美国西部;
D1=0 表示其他地区.



D3=1 位于美国南部;
D3=0 表示其他地区.



虚拟变量模型的构建规则

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。否则, 就会陷入“虚拟变量陷阱”。(为什么?)
- 规则3: 在虚拟变量的设置中: 基础类型、肯定类型取值为1; 比较类型、否定类型取值为0。

| 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?



虚拟变量模型的构建规则：示例

建模1：(正确模型)使用m-1个虚拟变量，并设定为有截距：

$$\text{Salary}_i = \beta_1 + \beta_2 D2_i + \beta_3 D3_i + u_i$$

建模2：(正确模型)使用m个虚拟变量，并设定为无截距：

$$\text{Salary}_i = \alpha_1 D1_i + \alpha_2 D2_i + \alpha_3 D3_i + u_i$$

建模3：(错误模型)使用m个虚拟变量，并设定为有截距：

$$\text{Salary}_i = \gamma_0 + \gamma_1 D1_i + \gamma_2 D2_i + \gamma_3 D3_i + u_i$$

提问：

- 模型1和模型2的回归系数涵义是一样的么？
- 执意采用OLS方法估计模型3，会有什么后果？



虚拟变量模型的构建规则：R软件示例

```
mod.main3 <- "Salary ~1+ D1+D2 +D3"  
lm.main3 <- lm(mod.main3, data_demon)  
summary(lm.main3)
```

Call:
lm(formula = mod.main3, data = data_demon)

Residuals:
Min 1Q Median 3Q Max
-14161 -4566 -1638 4632 15625

Coefficients: (1 not defined because of singularities)
Estimate Std. Error t value Pr(>|t|)

对于错误的建模3，有些统计软件（如R软件）会自动去掉一个多余的虚拟变量

$$Salary_i = \gamma_0 + \gamma_1 D1_i + \gamma_2 D2_i + \gamma_3 D3_i + u_i$$

4.2.4 同时含有一个定性变量 和定量变量的ANOVA模型



公立学校教师薪水案例

我们继续以公立学校教师薪水案例进行分析和说明。



数据

把定性变量 *Region* 进行数据变换，得到数据：

把定性变量 *Region* 处理成虚拟变量 n=51

state	Region	Salary	Spend	D1	D2	D3
Connecticut	M.E.N	60822	12436	0	1	0
Illinois	M.E.N	58246	9275	0	1	0
Indiana	M.E.N	47831	8935	0	1	0
Iowa	M.E.N	43130	7807	0	1	0
Kansas	M.E.N	43334	8373	0	1	0
Maine	M.E.N	41596	11285	0	1	0
Massachusetts	M.E.N	58624	12596	0	1	0
Michigan	M.E.N	54895	9880	0	1	0

Showing 1 to 8 of 51 entries

Previous

1

2

3

4

5

6

7

Next



有截距虚拟变量模型：PRM

我们可以构建薪水 ($Salary$) 对区域虚拟变量 ($D2; D3$) 和定量变量 $Spend$ 的有截距总体回归模型PRM:

$$Salary_i = \beta_1 + \beta_2 D2_i + \beta_3 D3_i + \lambda Spend_i + u_i$$

理论上，我们可以得到三个区域教师薪水的期望值：

$$E(Salary|D2 = 1, D3 = 0) = \beta_1 + \beta_2 + \lambda Spend \quad (\text{M.E.N})$$

$$E(Salary|D2 = 0, D3 = 1) = \beta_1 + \beta_3 + \lambda Spend \quad (\text{South})$$

$$E(Salary|D2 = 0, D3 = 0) = \beta_1 + \lambda Spend \quad (\text{West})$$

$D2 = 1$ 位于中/东/北部;
 $D2 = 0$ 表示其他地区.

北

中部

东

西

南

$D1=1$ 位于美国西部;
 $D1=0$ 表示其他地区.

$D3=1$ 位于美国南部;
 $D3=0$ 表示其他地区.



有截距虚拟变量模型：OLS估计

$$\widehat{\text{Salary}} = + 28694.92 - 2954.13D2 - 3112.19D3 + 2.34Spend$$

(t) (8.7953) (-1.5860) (-1.7101) (6.5152)
(se) (3262.5213) (1862.5756) (1819.8725) (0.3592)

(fitness) $R^2 = 0.4977; \bar{R}^2 = 0.4656$
 $F^* = 15.52; p = 0.0000$

提问1：大白话解释上述回归函数！

D2=1 位于中/东/北部;
D2=0 表示其他地区.

思考1：基准组是什么？谁是协变量？

北

思考2：三条线为什么是平行的？

西

中部

东

思考3：统计上来看，南部线和西部线

D1=1 位于美国西部;
D1=0 表示其他地区.

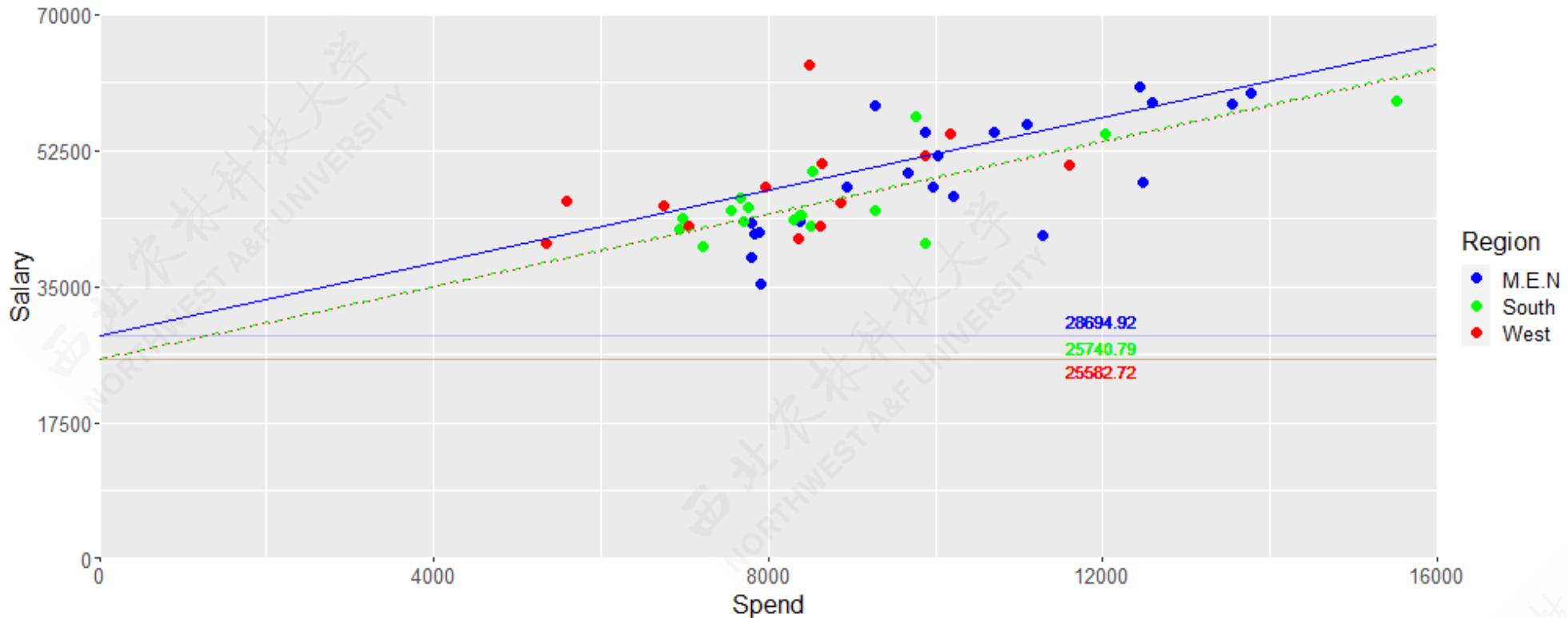
南

是不一样的么？

D3=1 位于美国南部;
D3=0 表示其他地区.



有截距虚拟变量模型：OLS估计





无截距虚拟变量模型：PRM2

我们可以构建薪水 (*Salary*) 对区域虚拟变量 ($D1; D2; D3$) 和定量变量 *Spend* 的无截距总体回归模型PRM:

$$Salary_i = \alpha_1 D1_i + \alpha_2 D2_i + \alpha_3 D3_i + \lambda Spend_i + u_i$$

理论上，我们可以得到三个区域教师薪水的期望值：

$$E(Salary|D1 = 0, D2 = 1, D3 = 0; Spend) = \alpha_2 + \lambda Spend$$

(M.E. 表示位于中/东/北部;
 $D2=0$ 表示其他地区.)

$$E(Salary|D1 = 0, D2 = 0, D3 = 1; Spend) = \alpha_3 + \lambda Spend$$

(South)

$$E(Salary|D1 = 1, D2 = 0, D3 = 0; Spend) = \alpha_1 + \lambda Spend$$

(West)

北

西

中部

东

南

D1=1 表示位于美国西部;
D1=0 表示其他地区.

D3=1 表示位于美国南部;
D3=0 表示其他地区.



无截距虚拟变量模型：OLS估计

$$\widehat{\text{Salary}} = + 28694.92D1 + 25740.79D2 + 25582.72D3 + 2.34Spend$$

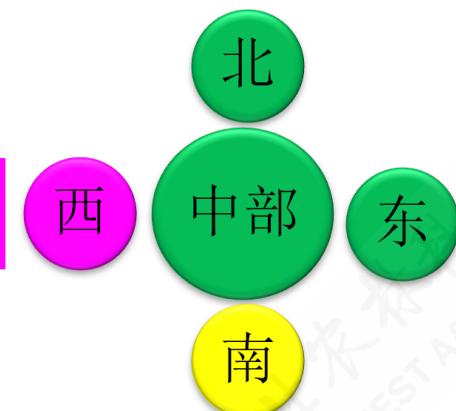
(t) (8.7953) (6.7627) (7.5372) (6.5152)
(se) (3262.5213) (3806.2835) (3394.1819) (0.3592)

(fitness) $R^2 = 0.9906$; $\bar{R}^2 = 0.9898$
 $F^* = 1235.97$; $p = 0.0000$

- 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？
- 思考4：有截距模型和无截距模型的图形为什么是一样的？

D2=1 位于中/东/北部;
D2=0 表示其他地区.

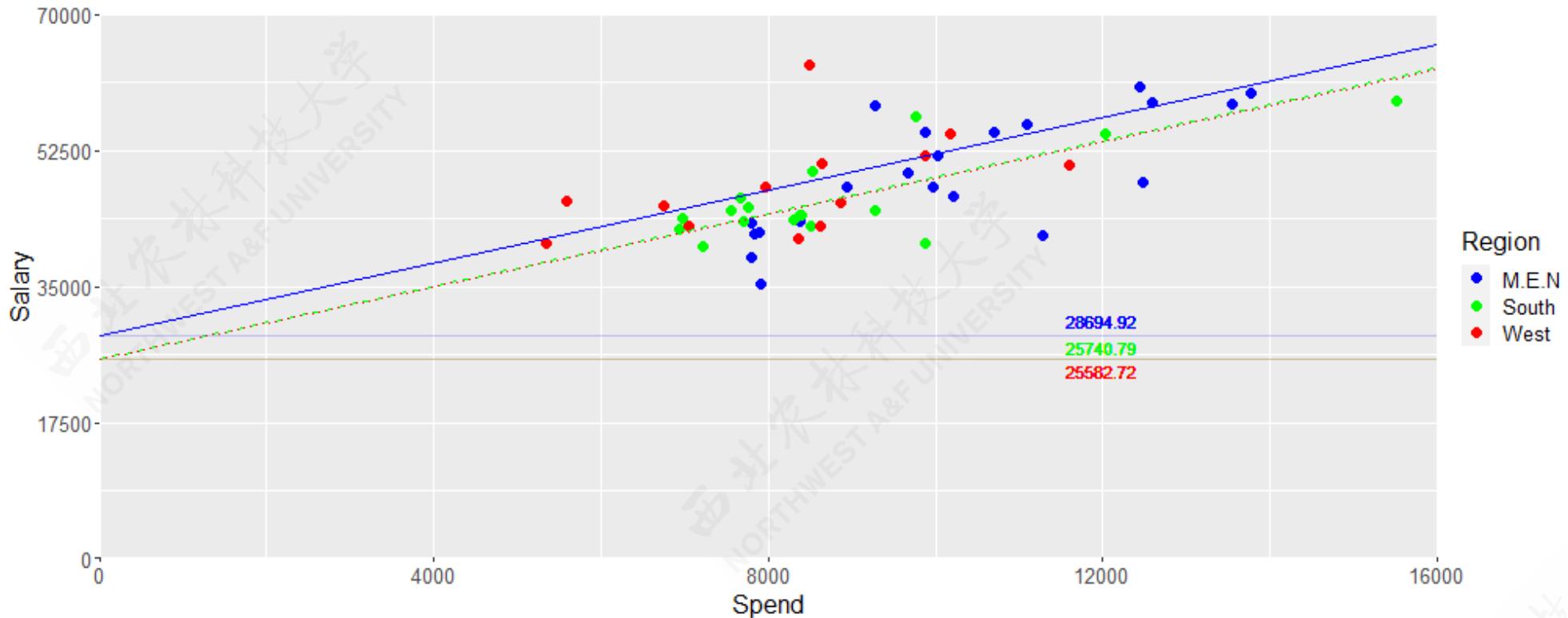
D1=1 位于美国西部;
D1=0 表示其他地区.



D3=1 位于美国南部;
D3=0 表示其他地区.



无截距虚拟变量模型：OLS估计



4.2.5 同时含有多个定性变量 和定量变量的ANOVA模型



虚拟变量的引入方式（定义）

如果自变量中存在k个定性变量 ($X_{1i}, X_{2i}, \dots, X_{ki}$)，而且每个定性变量还有自己的属性个数 ($X_{ki} (a_1, a_2, \dots, a_m)$)。那么把这些定性变量转换成各自的虚拟变量体系后，虚拟变量在模型中出现的关系则可以有多种形式：

- 虚拟变量以加法形式引入的回归模型：是指各个定性变量的虚拟变量体系，各自以独立项的形式出现在模型中。
- 虚拟变量以乘法形式引入的回归模型：是指各个定性变量的虚拟变量体系，存在相互以交叉项的形式出现在模型中。



虚拟变量的引入方式（定义）

如果自变量中存在k个定性变量 ($X_{1i}, X_{2i}, \dots, X_{ki}$)，而且每个定性变量还有自己的属性个数 ($X_{ki} (a_1, a_2, \dots, a_m)$)。那么把这些定性变量转换成各自的虚拟变量体系后，虚拟变量在模型中出现的关系则可以有多种形式：

- 虚拟变量以混合形式（既有加法形式也有乘法形式）引入的回归模型：是指各个定性变量的虚拟变量体系，及有各自以独立项的形式，也有相互以交叉项的形式出现在模型中。又具体分为两种情形：
 - 完全混合模型：两个定性变量的虚拟变量体系，既有各自独立项，又有它们相互间完全交叉项。
 - 部分混合模型：两个定性变量的虚拟变量体系，既有各自独立项，又有它们相互间不完全的交叉项（也即部分交叉）。



虚拟变量的引入方式（示例）

为了研究工人工资的影响因素，我们可以考虑如下变量：

- 定量变量：工资 $wage$; 年龄 age
- 定性变量：教育程度 $edu\{a_1 = \text{ill}, a_2 = \text{pri}, a_3 = \text{mid}, a_4 = \text{hig}\}$; 工作部门 $dpt\{a_1 = \text{tem}, a_2 = \text{per}\}$; 性别 $sex\{a_1 = \text{f}, a_2 = \text{m}\}$ 。

教育程度 edu : ill表示文盲；pri表示初等教育；mid表示中等教育；hig表示高等教育。

工作类型 dpt : tem表示临时工；per表示合同工。

性别 sex : f表示女性；m表示男性。



虚拟变量的引入方式（示例）

因为涉及到三个定性变量，我们可以将他们转换为各自的虚拟变量体系：

教育程度定性变量 edu ：

$$\text{edu}\{a_1 = \text{ill}, a_2 = \text{pri}, a_3 = \text{mid}, a_4 = \text{hig}\}$$

$$\text{dummy} \Rightarrow \begin{cases} \text{edu_ill} = \begin{cases} 1, & \text{ill} \\ 0, & \text{not ill} \end{cases} \\ \text{edu_pri} = \begin{cases} 1, & \text{pri} \\ 0, & \text{not pri} \end{cases} \\ \text{edu_mid} = \begin{cases} 1, & \text{mid} \\ 0, & \text{not mid} \end{cases} \\ \text{edu_hig} = \begin{cases} 1, & \text{hig} \\ 0, & \text{not hig} \end{cases} \end{cases}$$

工作类型定性变量 dpt ：

$$\text{dpt}\{a_1 = \text{tem}, a_2 = \text{per}\}$$

$$\text{dummy} \Rightarrow \begin{cases} \text{dpt_tem} = \begin{cases} 1, & \text{tem} \\ 0, & \text{not tem} \end{cases} \\ \text{dpt_per} = \begin{cases} 1, & \text{per} \\ 0, & \text{not per} \end{cases} \end{cases}$$

性别定性变量 sex ：

$$\text{sex}\{a_1 = \text{f}, a_2 = \text{m}\}$$

$$\text{dummy} \Rightarrow \begin{cases} \text{sex_f} = \begin{cases} 1, & \text{f} \\ 0, & \text{not f} \end{cases} \\ \text{sex_m} = \begin{cases} 1, & \text{m} \\ 0, & \text{not m} \end{cases} \end{cases}$$



虚拟变量的引入方式（示例：加法模型）

$$wage = \beta_1 + \beta_2 edu_{pri} + \beta_3 edu_{mid} + \beta_4 edu_{hig} + \beta_5 dpt_{per} + \beta_6 sex_m + \beta_7 age + u_i$$

假定我们关注这样两个群体：

A群体：年龄为30岁的、女性 ($sex_m = 0$)、受过高等教育 ($edu_{pri} = 0, edu_{mid} = 0, edu_{hig} = 1$) 的、拥有一份合同工的 ($dpt_{per} = 1$)。

$$\begin{aligned} E(wage | age = 30; & edu_{pri} = 0; edu_{mid} = 0; edu_{hig} = 1; dpt_{per} = 1; sex_m = 0) \\ &= \beta_1 + \beta_2(0) + \beta_3(0) + \beta_4(1) + \beta_5(1) + \beta_6(0) + \beta_7(30) \\ &= \beta_1 + \beta_4 + \beta_5 + 30\beta_7 \end{aligned}$$

B群体：年龄为30岁的、女性 ($sex_m = 0$)、受过高等教育 ($edu_{pri} = 0, edu_{mid} = 0, edu_{hig} = 1$) 的、拥有一份临时工的 ($dpt_{per} = 0$)。

$$\begin{aligned} E(wage | age = 30; & edu_{pri} = 0; edu_{mid} = 0; edu_{hig} = 1; dpt_{per} = 0; sex_m = 0) \\ &= \beta_1 + \beta_2(0) + \beta_3(0) + \beta_4(1) + \beta_5(0) + \beta_6(0) + \beta_7(30) \\ &= \beta_1 + \beta_4 + 30\beta_7 \end{aligned}$$



虚拟变量的引入方式（示例：加法模型）

$$wage = \beta_1 + \beta_2 edu_{pri} + \beta_3 edu_{mid} + \beta_4 edu_{hig} + \beta_5 dpt_{per} + \beta_6 sex_m + \beta_7 age + u_i$$

如果 $\beta_5 > 0$, 这将意味着:

- 只要拥有一份合同工 ($dpt_{per} = 1$)。那么, 在其他同等情况下, 这个人的工资都要高于拥有一份临时工 ($dpt_{per} = 0$) 的人。——无论是高学历的同等条件 ($edu_{pri} = 0, edu_{mid} = 0, edu_{hig} = 1$) , 还是中等学历的同等条件 ($edu_{pri} = 1, edu_{mid} = 0, edu_{hig} = 0$) , 还是文盲学历的同等条件 ($edu_{pri} = 0, edu_{mid} = 0, edu_{hig} = 0$) 。
- 换言之, 工作类型 (dpt) 与受教育程度 (edu) 是独立地作用于工资 (wage) 的!



虚拟变量的引入方式（示例：乘法模型）

$$\begin{aligned} wage = & + \beta_1 + \beta_2 sex_m + \beta_3 age + \beta_4 edu_{pri} * dpt_{per} \\ (\text{cont.}) & + \beta_5 dpt_{per} * edu_{mid} + \beta_6 dpt_{per} * edu_{hig} + u_i \end{aligned}$$

假定我们关注这样两个群体：

A群体： 年龄为30岁的、女性 ($sex_m = 0$)、受过高等教育 ($edu_{pri} = 0, edu_{mid} = 0, edu_{hig} = 1$) 的、拥有一份合同工的 ($dpt_{per} = 1$)。

$$\begin{aligned} E(wage | age = 30; edu_{pri} = 0; edu_{mid} = 0; edu_{hig} = 1; dpt_{per} = 1; sex_m = 0) \\ = & + \beta_1 + \beta_2(0) + \beta_3(30) + \beta_4(0) \cdot (1) + \beta_5(1) \cdot (0) + \beta_6(1) \cdot (1) \\ = & + \beta_1 + 30\beta_3 + \beta_6 \end{aligned}$$

B群体： 年龄为30岁的、女性 ($sex_m = 0$)、受过高等教育 ($edu_{pri} = 0, edu_{mid} = 0, edu_{hig} = 1$) 的、拥有一份临时工的 ($dpt_{per} = 0$)。



虚拟变量的引入方式（示例：乘法模型）

$$wage = + \beta_1 + \beta_2 sex_m + \beta_3 age + \beta_4 edu_{pri} * dpt_{per}$$

(cont.) $+ \beta_5 dpt_{per} * edu_{mid} + \beta_6 dpt_{per} * edu_{hig} + u_i$

如果 $\beta_6 > 0$ 且显著，这将意味着：

- 在其他同等情况下，一个拥有一份合同工 ($dpt_{per} = 1$) 且拥有高学历 ($edu_{hig} = 1$) 的人。这个人的工资都要高于拥有一份临时工 ($dpt_{per} = 0$) 或没有受过高学历教育 ($edu_{hig} = 0$) 的人。——包括：临时工&文盲；临时工&初等学历；临时工&中等学历；合同工&文盲；合同工&初等学历；合同工&中等学历。
- 换言之，工作类型 (dpt) 与受教育程度 (edu) 是交互地作用于工资 (wage) 的！
——此处重点针对拥有高学历还是不拥有高学历。



虚拟变量的引入方式（示例：混合模型）

$$wage = + \beta_1 + \beta_2 sex_m + \beta_3 dpt_{per} + \beta_4 age + \beta_5 edu_{pri} * dpt_{per}$$

(cont.) $+ \beta_6 edu_{mid} * dpt_{per} + \beta_7 edu_{hig} * dpt_{per} + u_i$

假定我们关注这样两个群体：

A群体： 年龄为30岁的、女性 ($sex_m = 0$)、受过高等教育 ($edu_{pri} = 0, edu_{mid} = 0, edu_{hig} = 1$) 的、拥有一份合同工的 ($dpt_{per} = 1$)。

$$\begin{aligned} E(wage | age = 30; & edu_{pri} = 0; edu_{mid} = 0; edu_{hig} = 1; dpt_{per} = 1; sex_m = 0) \\ &= + \beta_1 + \beta_2(0) + \beta_3(1) + \beta_4(30) + \beta_5(0) \cdot (1) + \beta_6(0) \cdot (1) + \beta_7(1) \cdot (1) \\ &= + \beta_1 + \beta_3 + 30\beta_4 + \beta_7 \end{aligned}$$

B群体： 年龄为30岁的、女性 ($sex_m = 0$)、受过高等教育 ($edu_{pri} = 0, edu_{mid} = 0, edu_{hig} = 1$) 的、拥有一份临时工的 ($dpt_{per} = 0$)。

4.2.6 印度工人工资案例



数据：原始

印度工人工资：114位印度工人工资方面的数据如下。

印度工人工资 (n=114)

wage	age	edu	dpt	sex
117	26	pri	per	f
375	42	pri	per	f
175	33	pri	per	f
100	33	pri	per	f
162.5	30	pri	per	f
300	35	mid	per	f
175	40	mid	per	f
287.5	50	mid	per	f

Showing 1 to 8 of 114 entries

Previous

1

2

3

4

5

...

15

Next



数据：变量定义

变量说明见下表：

变量定义及说明

variable	label	remark
obs	工人编号	序号(observations)
wage	工人工资	美元/周(\$/week)
age	年龄	岁(year)
edu	教育水平	ill=文盲(illiteracy); pri=初等教育(primary); mid=中等教育(middle); hig=高等教育(higher)
dpt	合同类型	tem=短期合同(temporary); per=长期合同(permanent)
sex	性别	f=女(female); m=男(male)



数据：定性变量的属性统计

定性变量及其属性的统计表($n=119$)

属性	频次
ill	74
pri	17
mid	17
hig	6
tem	72
per	42
f	89
m	25



数据：定性变量的虚拟变量变换

将基础组设定为{文盲，临时工，女性}（也即{illiteracy, temporary, female}）。则可以将全部定性变量的基础组属性{illiteracy, temporary, female}分别设置为虚拟变量 `edu_ill`、`dpt_tem` 和 `sex_f`。

教育程度定性变量 `edu`：

$$edu\{a_1 = ill, a_2 = pri, a_3 = mid, a_4 = hig\}$$

$$dummy \Rightarrow \begin{cases} edu_ill = \begin{cases} 1, & ill \\ 0, & not ill \end{cases} \\ edu_pri = \begin{cases} 1, & pri \\ 0, & not pri \end{cases} \\ edu_mid = \begin{cases} 1, & mid \\ 0, & not mid \end{cases} \\ edu_hig = \begin{cases} 1, & hig \\ 0, & not hig \end{cases} \end{cases}$$

工作类型定性变量 `dpt`：

$$dpt\{a_1 = tem, a_2 = per\}$$

$$dummy \Rightarrow \begin{cases} dpt_tem = \begin{cases} 1, & tem \\ 0, & not tem \end{cases} \\ dpt_per = \begin{cases} 1, & per \\ 0, & not per \end{cases} \end{cases}$$

性别定性变量 `sex`：

$$sex\{a_1 = f, a_2 = m\}$$

$$dummy \Rightarrow \begin{cases} sex_f = \begin{cases} 1, & f \\ 0, & not f \end{cases} \\ sex_m = \begin{cases} 1, & m \\ 0, & not m \end{cases} \end{cases}$$



数据：教育变量的虚拟变量变换

用虚拟变量系统完全表达定性变量edu (n=114)

edu	edu_ill	edu_pri	edu_mid	edu_hig
pri	0	1	0	0
pri	0	1	0	0
pri	0	1	0	0
pri	0	1	0	0
pri	0	1	0	0
mid	0	0	1	0
mid	0	0	1	0
mid	0	0	1	0

Showing 1 to 8 of 114 entries

Previous

1

2

3

4

5

...

15

Next

NORTHWEST A&F UNIVERSITY



数据：工作部门变量的虚拟变量变换

用虚拟变量系统完全表达定性变量dpt (n=114)

Showing 1 to 8 of 114 entries

Previous

1

2

3

15

Next



数据：性别变量的虚拟变量变换

用虚拟变量系统完全表达定性变量sex (n=114)

sex	sex_f	sex_m
f	1	0
f	1	0
f	1	0
f	1	0
f	1	0
f	1	0
f	1	0
f	1	0

Showing 1 to 8 of 114 entries

Previous

1

2

3

4

5

...

15

Next



加法模型：总体回归模型PRM

同时含虚拟变量和定量变量的、加法形式的经典回归模型：

$$wage = + \beta_1 + \beta_2 edu_{pri} + \beta_3 edu_{mid} + \beta_4 edu_{hig} + \beta_5 dpt_{per} + \beta_6 sex_m + \beta_7 age + u_i$$

OLS估计的简要报告如下：

$$\begin{aligned}\widehat{wage} = & + 6.79 & + 23.96 edu_{pri} & + 61.59 edu_{mid} & + 150.49 edu_{hig} \\(t) & (0.2130) & (0.7734) & (1.9867) & (3.0054) \\(se) & (31.8931) & (30.9789) & (31.0035) & (50.0725) \\(\text{cont.}) & + 31.16 dpt_{per} & - 83.20 sex_m & + 3.99 age \\(t) & (1.3141) & (-3.0819) & (4.5129) \\(se) & (23.7120) & (26.9981) & (0.8835) \\(\text{fitness}) R^2 & = 0.3450; \bar{R}^2 & = 0.3083 \\F^* & = 9.39; \quad p = 0.0000\end{aligned}$$



加法模型 : EViews 报告

Dependent Variable: WAGE

Method: Least Squares

Date: Time:

Sample: 1 114

Included observations: 114

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	6.793986	31.89314	0.213023	0.8317
EDU_PRI	23.96067	30.97895	0.773450	0.4410
EDU_MID	61.59460	31.00348	1.986699	0.0495
EDU_HIG	150.4891	50.07253	3.005423	0.0033
DPT_PER	31.16083	23.71204	1.314136	0.1916
SEX_M	-83.20477	26.99814	-3.081871	0.0026
AGE	3.987214	0.883512	4.512911	0.0000
R-squared	0.345032	Mean dependent var	146.4085	
Adjusted R-squared	0.308305	S.D. dependent var	137.0137	
S.E. of regression	113.9518	Akaike info criterion	12.36887	
Sum squared resid	1389397.	Schwarz criterion	12.53688	
Log likelihood	-698.0254	Hannan-Quinn criter.	12.43705	
F-statistic	9.394456	Durbin-Watson stat	1.949593	
Prob(F-statistic)	0.000000			



加法模型：基础组

- 基础组（文盲 & 短期合同 & 女性），也即(**illiteracy & temporary & female**)的期望工资收入为（给定年龄为30岁）：

$$\begin{aligned}E(wage | age = 30; edu_{pri} = 0; edu_{mid} = 0; edu_{hig} = 0; dpt_{per} = 0; sex_m = 0) \\= + \beta_1 + \beta_2(0) + \beta_3(0) + \beta_4(0) + \beta_5(0) + \beta_6(0) + \beta_7(30) \\= + \beta_1 + 30\beta_7\end{aligned}$$

$$\begin{aligned}(\widehat{wage} | age = 30; edu_{pri} = 0; edu_{mid} = 0; edu_{hig} = 0; dpt_{per} = 0; sex_m = 0) \\= + \hat{\beta}_1 + \hat{\beta}_2(0) + \hat{\beta}_3(0) + \hat{\beta}_4(0) \\+ \hat{\beta}_5(0) + \hat{\beta}_6(0) + \hat{\beta}_7(30) \\= + [6.79] + [23.96] \cdot (0) + [61.59] \cdot (0) + [150.49] \cdot (0) \\+ [31.16] \cdot (0) + [-83.20] \cdot (0) + [3.99] \cdot (30) \\= 126.4104\end{aligned}$$



加法模型：比较组1

比较组1（高等学历 & 短期合同 & 女性），也即(**high & temporary & female**)的期望工资收入为（给定年龄为30岁）：

$$\begin{aligned} E(wage | age = 30; edu_{pri} = 0; edu_{mid} = 0; edu_{hig} = 1; dpt_{per} = 0; sex_m = 0) \\ = + \beta_1 + \beta_2(0) + \beta_3(0) + \beta_4(1) + \beta_5(0) + \beta_6(0) + \beta_7(30) \\ = + \beta_1 + \beta_4 + 30\beta_7 \end{aligned}$$

$$\begin{aligned} (\widehat{wage} | age = 30; edu_{pri} = 0; edu_{mid} = 0; edu_{hig} = 1; dpt_{per} = 0; sex_m = 0) \\ = + \hat{\beta}_1 + \hat{\beta}_2(0) + \hat{\beta}_3(0) + \hat{\beta}_4(1) \\ + \hat{\beta}_5(0) + \hat{\beta}_6(0) + \hat{\beta}_7(30) \\ = + [6.79] + [23.96] \cdot (0) + [61.59] \cdot (0) + [150.49] \cdot (1) \\ + [31.16] \cdot (0) + [-83.20] \cdot (0) + [3.99] \cdot (30) \\ = 276.8995 \end{aligned}$$



加法模型：比较组2

比较组2（高等教育 & 长期合同 & 女性），也即(**higher & permanent & female**)的期望工资收入为（给定年龄为30岁）：

$$\begin{aligned} E(wage | age = 30; edu_{pri} = 0; edu_{mid} = 0; edu_{hig} = 1; dpt_{per} = 1; sex_m = 0) \\ = + \beta_1 + \beta_2(0) + \beta_3(0) + \beta_4(1) + \beta_5(1) + \beta_6(0) + \beta_7(30) \\ = + \beta_1 + \beta_4 + \beta_5 + 30\beta_7 \end{aligned}$$

$$\begin{aligned} (\widehat{wage} | age = 30; edu_{pri} = 0; edu_{mid} = 0; edu_{hig} = 1; dpt_{per} = 1; sex_m = 0) \\ = + \hat{\beta}_1 + \hat{\beta}_2(0) + \hat{\beta}_3(0) + \hat{\beta}_4(1) \\ + \hat{\beta}_5(1) + \hat{\beta}_6(0) + \hat{\beta}_7(30) \\ = + [6.79] + [23.96] \cdot (0) + [61.59] \cdot (0) + [150.49] \cdot (1) \\ + [31.16] \cdot (1) + [-83.20] \cdot (0) + [3.99] \cdot (30) \\ = 308.0604 \end{aligned}$$



乘法模型：总体回归模型PRM

同时含虚拟变量和定量变量的、加法形式的经典回归模型：

$$wage = \beta_1 + \beta_2 sex_m + \beta_3 age + \beta_4 edu_{pri} * dpt_{per} + \beta_5 dpt_{per} * edu_{mid} + \beta_6 dpt_{per} * edu_{hig} + u_i$$

OLS估计的简要报告如下：

\widehat{wage}	$+ 22.92$	$- 71.41 sex_m$	$+ 4.11 age$	$+ 28.21 edu_{pri} * dpt_{per}$
(t)	(0.7366)	(-2.5720)	(4.6325)	(0.5166)
(se)	(31.1120)	(27.7639)	(0.8870)	(54.5979)
(cont.)	$+ 112.68 dpt_{per} * edu_{mid} + 33.12 dpt_{per} * edu_{hig}$			
(t)	(2.4012)	(0.5965)		
(se)	(46.9258)	(55.5274)		
(fitness)	$R^2 = 0.2872;$	$\bar{R}^2 = 0.2542$		
	$F^* = 8.70;$	$p = 0.0000$		



乘法模型 : EViews 报告

Dependent Variable: WAGE

Method: Least Squares

Date: Time:

Sample: 1 114

Included observations: 114

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	22.91756	31.11195	0.736616	0.4630
SEX_M	-71.40897	27.76392	-2.572006	0.0115
AGE	4.109012	0.887004	4.632462	0.0000
EDU_PRI*DPT_PER	28.20685	54.59786	0.516629	0.6065
EDU_MID*DPT_PER	112.6797	46.92583	2.401229	0.0180
EDU_HIG*DPT_PER	33.12241	55.52738	0.596506	0.5521
R-squared	0.287210	Mean dependent var	146.4085	
Adjusted R-squared	0.254211	S.D. dependent var	137.0137	
S.E. of regression	118.3237	Akaike info criterion	12.43592	
Sum squared resid	1512055.	Schwarz criterion	12.57993	
Log likelihood	-702.8476	Hannan-Quinn criter.	12.49437	
F-statistic	8.703470	Durbin-Watson stat	1.990210	
Prob(F-statistic)	0.000001			



乘法模型：基础组

- 基础组（文盲 & 短期合同 & 女性），也即(**illiteracy & temporary & female**)的期望工资收入为（给定年龄为30岁）：

$$\begin{aligned}E(wage | age = 30; edu_{pri} = 0; edu_{mid} = 0; edu_{hig} = 0; dpt_{per} = 0; sex_m = 0) \\= + \beta_1 + \beta_2(0) + \beta_3(30) + \beta_4(0) \cdot (0) + \beta_5(0) \cdot (0) + \beta_6(0) \cdot (0) \\= + \beta_1 + 30\beta_3\end{aligned}$$

$$\begin{aligned}(\widehat{wage} | age = 30; edu_{pri} = 0; edu_{mid} = 0; edu_{hig} = 0; dpt_{per} = 0; sex_m = 0) \\= + \hat{\beta}_1 + \hat{\beta}_2(0) + \hat{\beta}_3(30) + \hat{\beta}_4(0) \\+ \hat{\beta}_5(0) + \hat{\beta}_6(0) \\= + [22.92] + [-71.41] \cdot (0) + [4.11] \cdot (30) + [28.21] \cdot (0) \\+ [112.68] \cdot (0) + [33.12] \cdot (0) \\= 146.1879\end{aligned}$$



乘法模型：比较组1

比较组1（高等学历 & 短期合同 & 女性），也即(**high & temporary & female**)的期望工资收入为（给定年龄为30岁）：

$$\begin{aligned} E(wage | age = 30; edu_{pri} = 0; edu_{mid} = 0; edu_{hig} = 1; dpt_{per} = 0; sex_m = 0) \\ = + \beta_1 + \beta_2(0) + \beta_3(30) + \beta_4(0) \cdot (0) + \beta_5(0) \cdot (0) + \beta_6(0) \cdot (1) \\ = + \beta_1 + 30\beta_3 \end{aligned}$$

$$\begin{aligned} (\widehat{wage} | age = 30; edu_{pri} = 0; edu_{mid} = 0; edu_{hig} = 1; dpt_{per} = 0; sex_m = 0) \\ = + \hat{\beta}_1 + \hat{\beta}_2(0) + \hat{\beta}_3(30) + \hat{\beta}_4(0) \\ + \hat{\beta}_5(0) + \hat{\beta}_6(0) \\ = + [22.92] + [-71.41] \cdot (0) + [4.11] \cdot (30) + [28.21] \cdot (0) \\ + [112.68] \cdot (0) + [33.12] \cdot (0) \\ = 146.1879 \end{aligned}$$



乘法模型：比较组2

比较组2（高等教育 & 长期合同 & 女性），也即(**higher & permanent & female**)的期望工资收入为（给定年龄为30岁）：

$$\begin{aligned} E(wage | age = 30; edu_{pri} = 0; edu_{mid} = 0; edu_{hig} = 1; dpt_{per} = 1; sex_m = 0) \\ = + \beta_1 + \beta_2(0) + \beta_3(30) + \beta_4(0) \cdot (1) + \beta_5(1) \cdot (0) + \beta_6(1) \cdot (1) \\ = + \beta_1 + 30\beta_3 + \beta_6 \end{aligned}$$

$$\begin{aligned} (\widehat{wage} | age = 30; edu_{pri} = 0; edu_{mid} = 0; edu_{hig} = 1; dpt_{per} = 1; sex_m = 0) \\ = + \hat{\beta}_1 + \hat{\beta}_2(0) + \hat{\beta}_3(30) + \hat{\beta}_4(0) \\ + \hat{\beta}_5(0) + \hat{\beta}_6(1) \\ = + [22.92] + [-71.41] \cdot (0) + [4.11] \cdot (30) + [28.21] \cdot (0) \\ + [112.68] \cdot (0) + [33.12] \cdot (1) \\ = 179.3103 \end{aligned}$$

4.2.7 时间序列季节虚拟变量模型



时间序列季节虚拟变量模型

时间序列季节虚拟变量模型：是指时间变量以虚拟变量形式进入回归方程的模型，它是虚拟变量回归模型的一种特定形式及应用。

- 季节模式 (seasonal pattern)：大多数时间序列经济变量，通常表现出来的季节性往复行为或现象。
- 季节调整 (seasonal adjusted)：将时间序列经济变量的季节性变化成分去除，从而得到一个新的变量序列的处理过程。

事实上，一个时间序列经济变量往往同时存在四个成分，分别是：

- 季节成分 (seasonal component)
- 周期成分 (cyclical component)
- 趋势成分 (trend component)
- 严格随机成分 (strictly random component)



时间序列季节虚拟变量模型

如果把定性变量季节 (season: Q1; Q2; Q3; Q4) 变换为虚拟变量体系，则分别可以构建以第一季度为基础组的时间序列季节虚拟变量模型和无基础组的时间序列季节虚拟变量模型 (X_t 为定量变量)。

$$Y_t = \beta_1 + \beta_2 X_t + \lambda_2 D_{2t} + \lambda_3 D_{3t} + \lambda_4 D_{4t} + u_t$$
$$Y_t = \beta_2 X_t + \lambda_1 D_{1t} + \lambda_2 D_{2t} + \lambda_3 D_{3t} + \lambda_4 D_{4t} + u_t$$



交通事故案例



数据

交通事故数据 (n=108)

year	month	totacc
1981	1	40511
1981	2	36034
1981	3	40328
1981	4	37699
1981	5	38816
1981	6	38900
1981	7	38625
1981	8	39539

Showing 1 to 8 of 108 entries

Previous

1

2

3

4

5

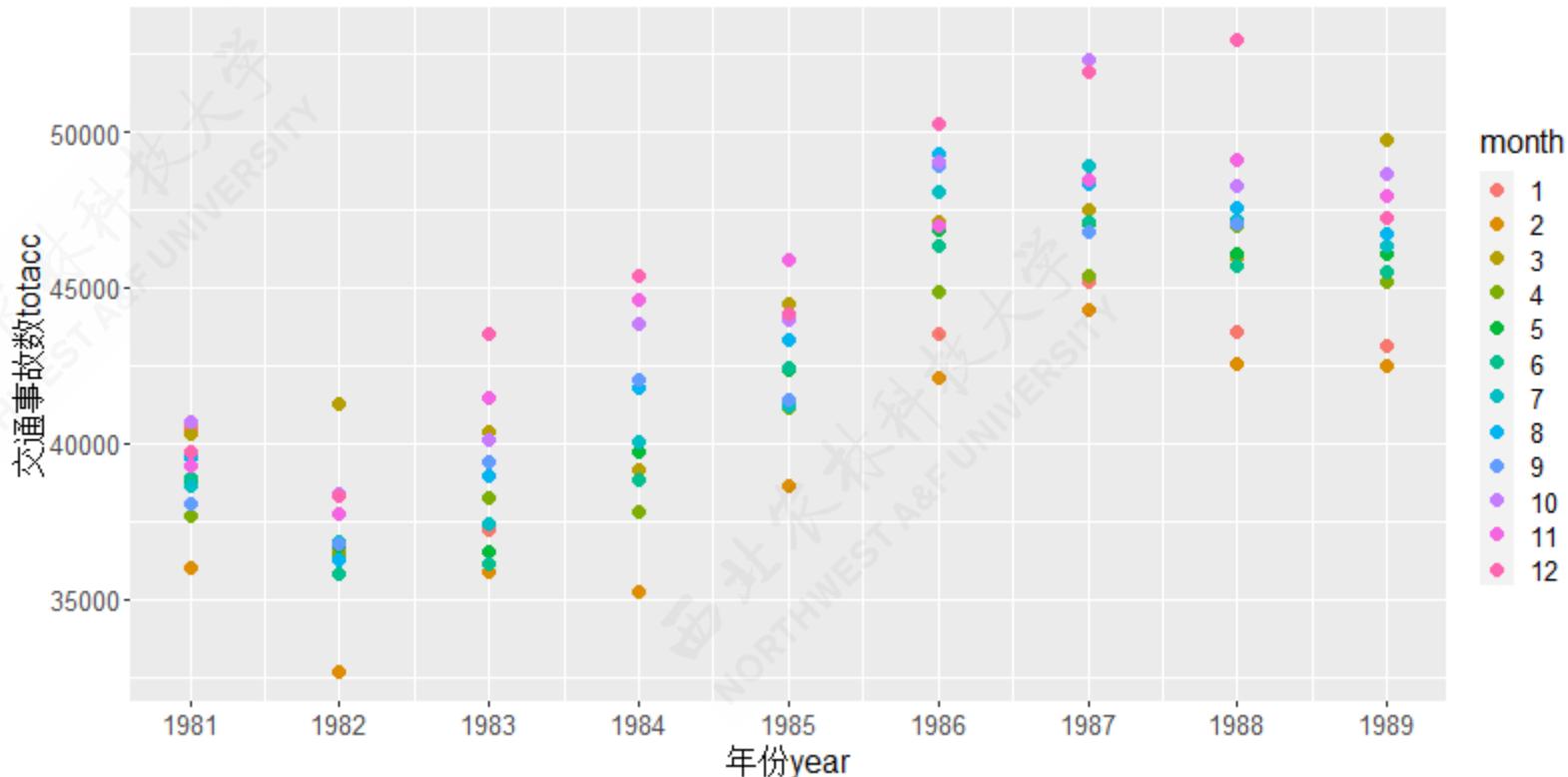
...

14

Next



散点图





虚拟变量

将月份month转换成虚拟变量系统($n=108$)

year	month	totacc	jan	feb	mar	apr	may	jun	jul	aug	sep	0
1981	1	40511	1	0	0	0	0	0	0	0	0	0
1981	2	36034	0	1	0	0	0	0	0	0	0	0
1981	3	40328	0	0	1	0	0	0	0	0	0	0
1981	4	37699	0	0	0	1	0	0	0	0	0	0
1981	5	38816	0	0	0	0	1	0	0	0	0	0
1981	6	38900	0	0	0	0	0	1	0	0	0	0
1981	7	38625	0	0	0	0	0	0	1	0	0	0
1981	8	39539	0	0	0	0	0	0	0	1	0	0

Showing 1 to 8 of 108 entries

Previous

1

2

3

4

5

...

14

Next



模型设定PRM

我们以1月份为基础组，构建如下的有截距虚拟变量回归模型：

$$\begin{aligned} \log(totacc) = & +\beta_1 + \beta_2 feb + \beta_3 mar + \beta_4 apr \\ (\text{cont.}) & + \beta_5 may + \beta_6 jun + \beta_7 aug + \beta_8 sep \\ (\text{cont.}) & + \beta_9 oct + \beta_{10} nov + \beta_{11} dec + u_i \end{aligned}$$



OLS估计

以上模型的OLS估计结果如下：

$\widehat{\log(totacc)}$	$= + 10.63$	$- 0.07feb$	$+ 0.06mar$	$- 0.00apr$
(t)	(433.2562)	(-1.5759)	(1.3715)	(-0.0072)
(se)	(0.0245)	(0.0425)	(0.0425)	(0.0425)
(cont.)	$+ 0.02may$	$+ 0.01jun$	$+ 0.05aug$	$+ 0.04sep$
(t)	(0.3778)	(0.1622)	(1.0867)	(0.8778)
(se)	(0.0425)	(0.0425)	(0.0425)	(0.0425)
(cont.)	$+ 0.08oct$	$+ 0.07nov$	$+ 0.10dec$	
(t)	(1.8779)	(1.6876)	(2.3376)	
(se)	(0.0425)	(0.0425)	(0.0425)	
(fitness)	$R^2 = 0.1644; \bar{R}^2 = 0.0783$			
	$F^* = 1.91;$	$p = 0.0529$		

4.2.8 分段线性回归模型



分段线性回归模型

分段现象：在经济关系中，当解释变量 X 的值达到某一水平/阀值 X^* 之前，与被解释变量之间存在某种线性关系；当解释变量 X 的值达到或者超过水平/阀值 X^* 以后，与被解释变量的关系就会发生变化。因而总体看来，似乎被明显“分段”了。

分段线性回归模型 (piecewise linear regression)：是指用虚拟变量估计不同水平/阀值的解释变量 X 对被解释变量 Y 的影响的一类线性回归模型。它是虚拟变量回归模型的一种特定形式及应用。

- 一个阀值的分段线性回归模型：

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

- 两个阀值的分段线性回归模型：

$$Y_i = \beta_1 + \beta_2 X_i + \lambda_1(X_i - X_1^*) D1_i + \lambda_2(X_i - X_2^*) D2_i + u_i$$



供给和需求分析案例



数据

供给和需求数据 ($n=54$)

demand	offer
1155	39.3
362	23.5
357	22.4
111	6.1
703	35.9
494	35.5
410	23.2
63	9.1

Showing 1 to 8 of 54 entries

Previous

1

2

3

4

5

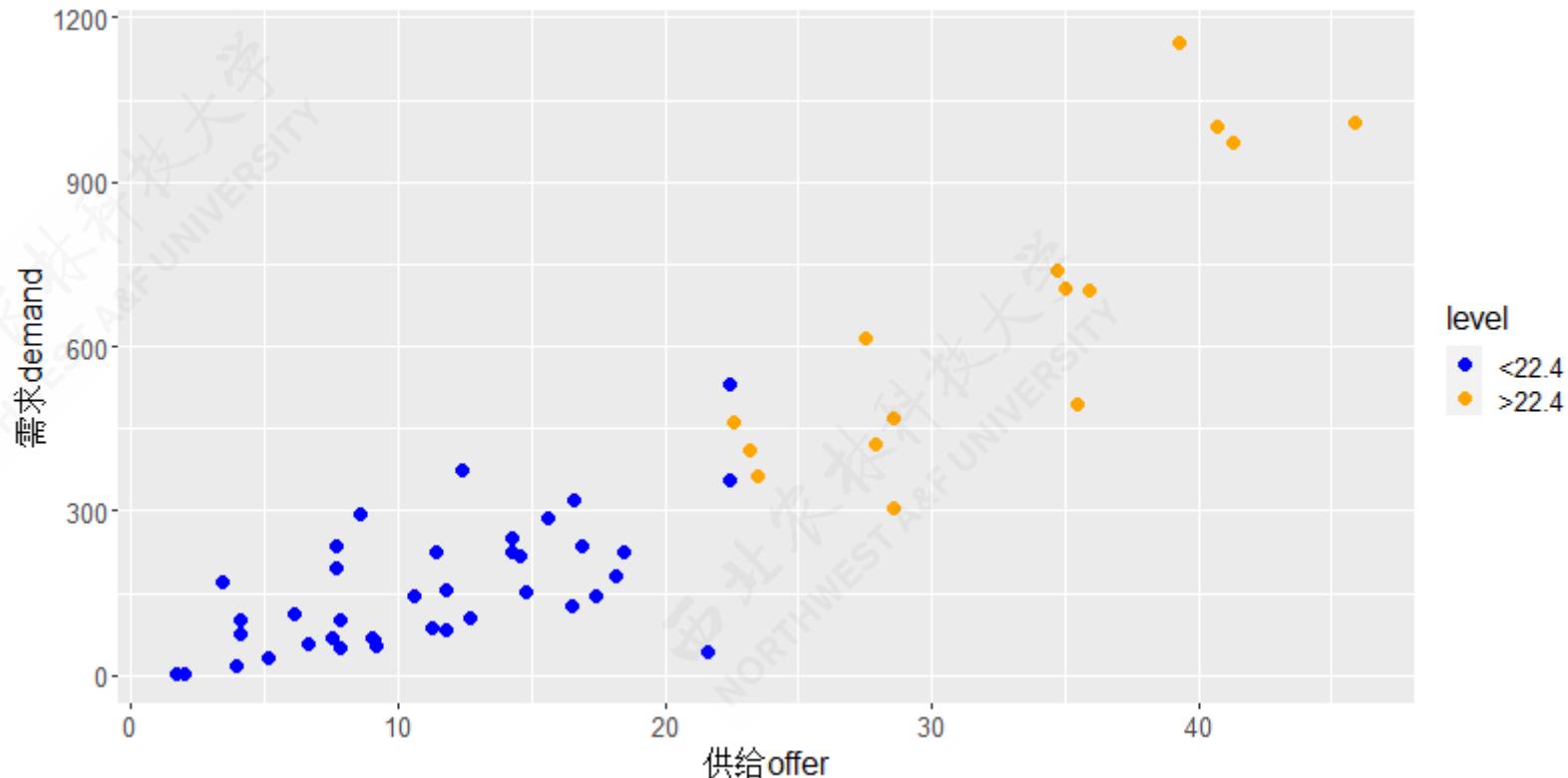
6

7

Next



散点图





虚拟变量体系

使用某些特定的方法，我们可以把数据的阀值设置为22.4，并对数据进行虚拟变量变换：

$$level \quad \{a_1 : < 22.4, a_2 : > 22.4\}$$
$$dummy \Rightarrow \begin{cases} D1 = \begin{cases} 1, & \text{offer} < 22.4 \\ 0, & \text{other} \end{cases} \\ D2 = \begin{cases} 1, & \text{offer} > 22.4 \\ 0, & \text{other} \end{cases} \end{cases}$$



虚拟变量变换

用offer的阈值22.4来设定虚拟变量系统($n=54$)

demand	offer	D1	D2
1155	39.3	0	1
362	23.5	0	1
357	22.4	0	0
111	6.1	1	0
703	35.9	0	1
494	35.5	0	1
410	23.2	0	1
63	9.1	1	0

Showing 1 to 8 of 54 entries

Previous

1

2

3

4

5

6

7

Next



模型估计

我们以offer=22.4为门阀值，构建如下的分段线性回归模型：

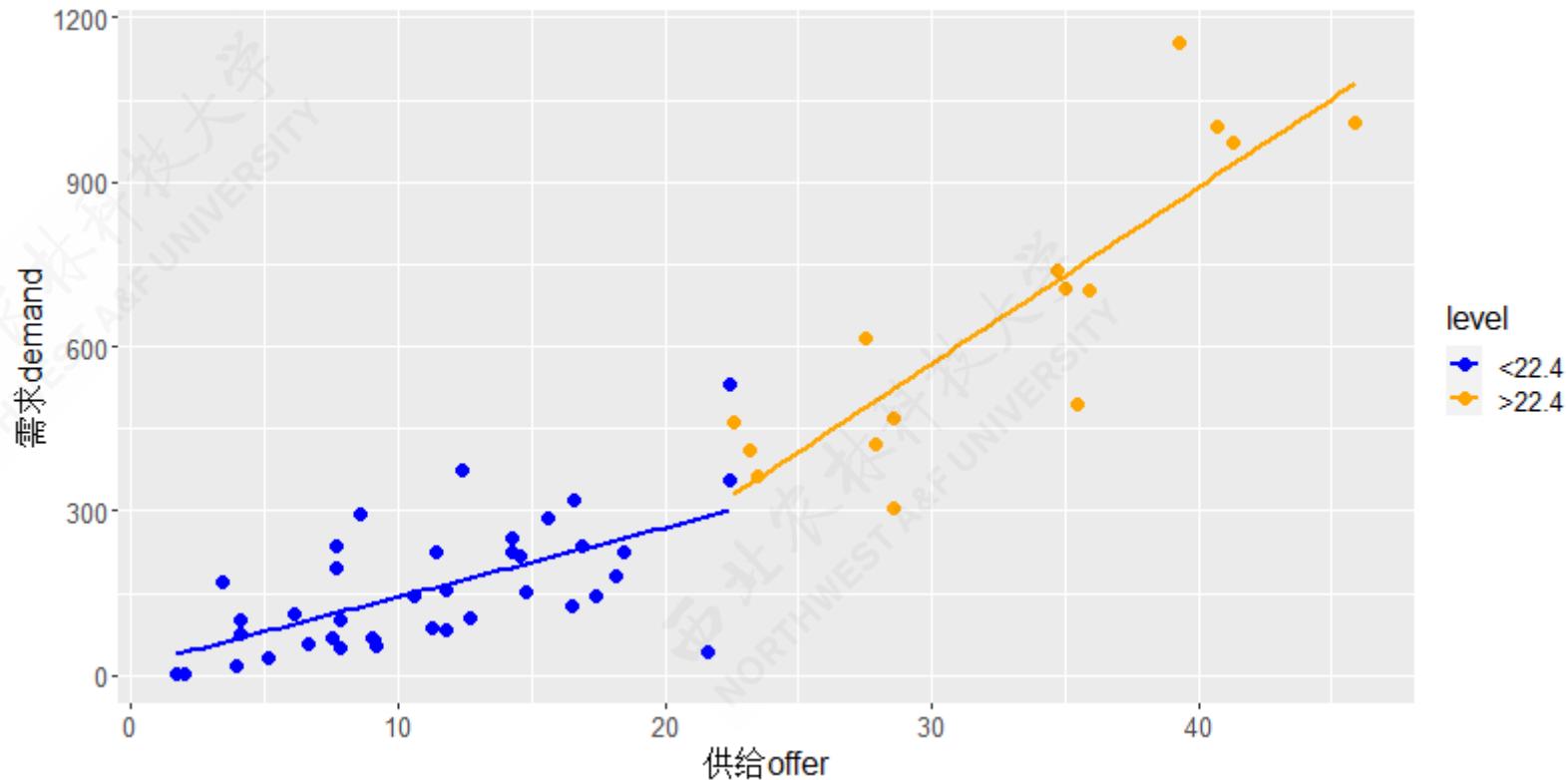
$$demand = \beta_1 + \beta_2 offer + \beta_3 I(offer - 22.4) * D1 + u_i$$

以上模型的OLS估计结果如下：

$$\begin{array}{cccc} \widehat{demand} = & -434.61 & +33.19offer & -19.88I(offer - 22.4) * D1 \\ (t) & (-4.8602) & (11.1818) & (-4.0721) \\ (se) & (89.4227) & (2.9685) & (4.8828) \\ (\text{fitness}) & R^2 = 0.8630; \bar{R}^2 = 0.8576 & & \\ & F^* = 160.60; p = 0.0000 & & \end{array}$$



回归线



本章结束

