

计量经济学 Eviews 实验指导书

Lab 8 虚拟变量回归模型

胡华平

2018/5/4

目录

虚拟变量回归模型

实验目的及要求

- 目的：掌握虚拟变量模型的设置和分析方法。
- 要求：熟悉虚拟变量的设置方法；理解定性变量和定量变量模型的内涵；熟练加法模型和乘法模型的运用原理。

实验原理

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。例如，性别变量的具体取值往往为男或女。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。

Definition 1 (定量变量). 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。

Definition 2 (定性变量). 定性变量（**Qualitative variable**）一般也称为分类变量，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。

Definition 3 (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。

如何把定性变量转换为虚拟变量？

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。实际上，我们可以用一套虚拟变量体系来完全表达一个定性变量。然后按照一定的规则构建虚拟变量回归模型，从而避免参数估计、模型解释等问题的出现。

Definition 4 (虚拟变量). 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 ：

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

Definition 5 (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：

$$X\{a_1, a_2, \dots, a_m\} \Rightarrow \begin{cases} D_1 = \begin{cases} 1, & \text{if } a_1 \\ 0, & \text{if not } a_1 \end{cases} \\ \vdots \\ D_k = \begin{cases} 1, & \text{if } a_k \\ 0, & \text{if not } a_k \end{cases} \\ \vdots \\ D_m = \begin{cases} 1, & \text{if } a_m \\ 0, & \text{if not } a_m \end{cases} \end{cases} \quad (2)$$

例如,定性变量肤色(X)具有3个属性($m = 3$),具体为 $X\{a_1 = \text{yellow}, a_2 = \text{white}, a_3 = \text{black}\}$,则可以构造出如下的虚拟变量体系¹:

$$X\{a_1 = \text{yellow}, a_2 = \text{white}, a_3 = \text{black}\} \quad (3)$$

$$\Rightarrow \begin{cases} D_1 = \begin{cases} 1, & \text{yellow} \\ 0, & \text{not yellow} \end{cases} \\ D_2 = \begin{cases} 1, & \text{white} \\ 0, & \text{not white} \end{cases} \\ D_3 = \begin{cases} 1, & \text{black} \\ 0, & \text{not black} \end{cases} \end{cases} \quad (4)$$

如何理解虚拟变量回归模型？

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）²。

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

根据虚拟变量引入模型方式的不同，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

¹一个定性变量如果有 m 个属性，那么可以用 m 个虚拟变量完全表达该定性变量，也可以用 $(m - 1)$ 个虚拟变量充分表达该定性变量。

²方差分析模型（Analysis of variance, ANOVA）常用来分析量化的因变量 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

- 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³
 - 完全混合模型
 - 部分混合模型

根据虚拟变量模型是否参照基础组，可以划分为⁴：

- 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。

根据模型中的因变量 Y 是否取对数，可以划分为⁵：

- 经典线性模型：因变量为 Y
- 半对数模型：因变量为 $\ln(Y)$

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（具体请参看节??）
- 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（具体请参看节??）

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。显然，不同类型的虚拟变量模型设置，具有不同的经济学含义。甚至回归方程系数解读的直观性，模型构建意图表达的直接性等，也存在较大差异，都需要对各种备选的、可行的模型进行反复测试和甄选。

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（具体回归方程设置见节??和节??）：

- 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
- ...

实验内容

1. 采用最小二乘法建立主回归模型
2. 自相关问题模型的侦察方法

³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），但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）。

⁴如果理论要求与基础组对比，则理论模型必须设置为有截距回归模型；否则，理论模型需要设置无截距回归模型。

⁵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。

- a. 残差序列观察法 (描点图法): 绘制 e_t 序列的描点图 (dot plot)
 - b. 残差序列观察法 (描点图法): 确定滞后阶数并分别绘制 e_t 序列与 e_{t-1}, e_{t-2}, \dots 序列的散点图 (scatter plot)
 - c. 辅助回归法: 构建残差 e_t 序列对 e_{t-1}, e_{t-2}, \dots 序列的辅助回归方程
 - d. 自相关和偏相关分析法: Eviews 菜单操作对残差 e_t 序列进行自相关和偏相关分析 (注意滞后阶数的选择)
 - e. Durbin-Watson 检验法: 分析 Eviews 报告中的 D-W 统计量
 - f. 拉格朗日检验法 (LM-test): Eviews 菜单操作进行布罗施-戈弗雷 (Breusch-Goldfrey) 的拉格朗日检验 (B-G LM test)
3. 自相关问题模型的矫正方法:
- a. 广义最小二乘法 (GLS): 一阶差分法变换
 - b. 广义最小二乘法 (GLS): 基于残差辅助方程近似得到 ρ
 - c. 广义最小二乘法 (GLS): 基于 D-W 统计量近似计算得到 ρ
 - d. 广义最小二乘法 (GLS): 基迭代法近似计算得到 ρ
 - e. 一致标准误校正法 (HAC): 尼威-威斯特 (Newey-West) 校正法

实验案例——印度工人工资

印度工人工资: 表??给出给出了 114 位印度工人在 wage 工人工资, age 年龄, edu 教育水平, dpt 合同类型, sex 性别等方面的数据。

变量说明见表??:

主要实验步骤

导入数据并进行预处理

- 目标:
- 思路:
- 新建 Eviews 工作文件 (见图??)
 - 提示: Excel 数据, 每个同学的 Y 数据都不同, 找到自己学号对应下的 Y
 - Eviews 菜单操作:
 - a. 依次操作: File⇒New⇒Workfile
 - b. 进行 workfile create 引导设置:
 - * workfile structure type: unstructured/undatede
 - * data range: 114

表 1: 印度工人工资 (n=114)

obs	wage	age	edu	dpt	sex
1	117.00	26	primary	permanent	female
2	375.00	42	primary	permanent	female
3	175.00	33	primary	permanent	female
4	100.00	33	primary	permanent	female
5	162.50	30	primary	permanent	female
110	25.00	18	illiteracy	temporary	male
111	25.00	11	illiteracy	temporary	male
112	75.00	45	illiteracy	temporary	male
113	53.84	14	illiteracy	temporary	male
114	50.00	26	illiteracy	temporary	male

表 2: 变量定义及说明

variable	label	remark
obs	工人编号	序号
wage	工人工资	美元/周
age	年龄	岁
edu	教育水平	illiteracy= 文盲; primary= 初等教育; secondary= 中等教育; higher= 高等教育
dpt	合同类型	temporary= 短期合同; permanent= 长期合同
sex	性别	female= 女; male= 男

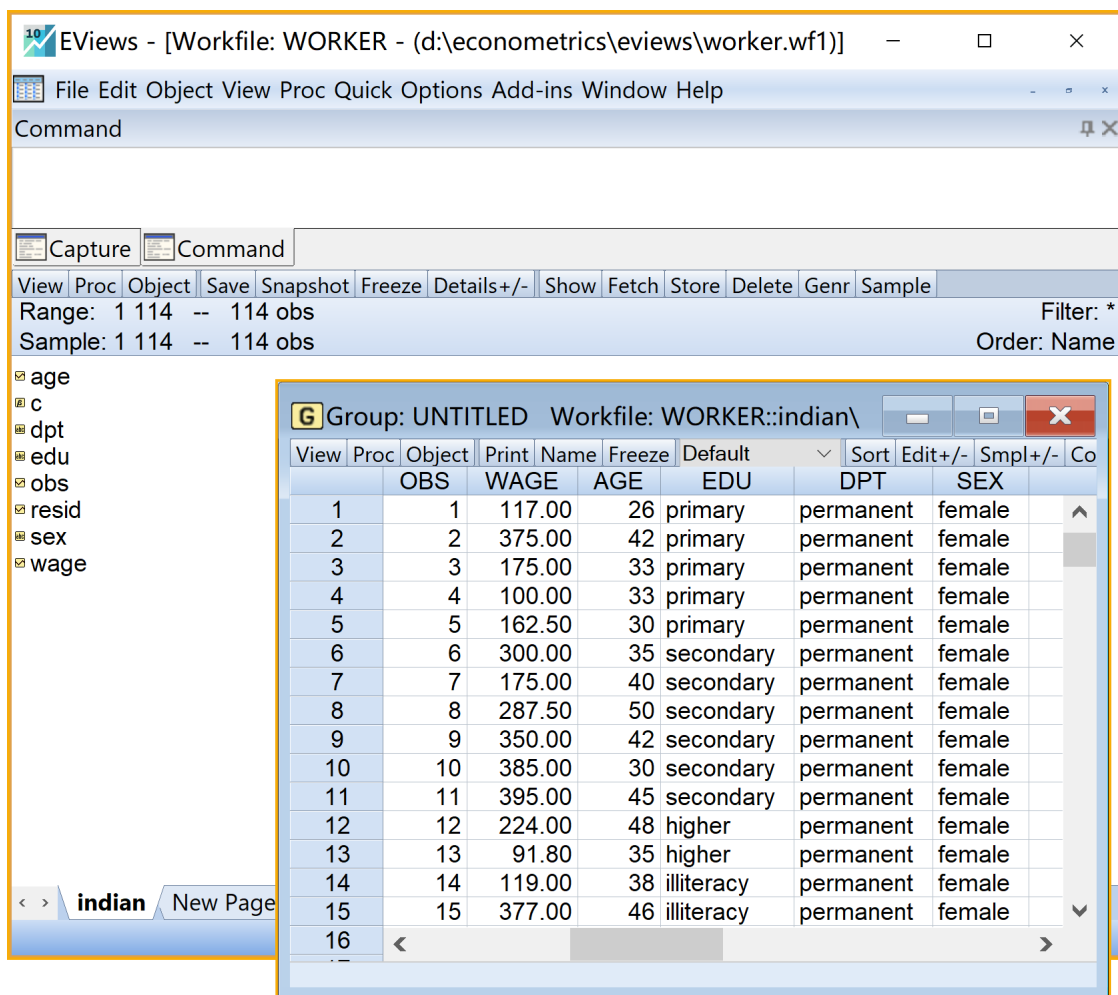


图 1: 导入数据的 Eviews 视窗

- * workfile names(optional):
 - WF: worker (建议命名)
 - Page: indian (建议命名)
- Eviews 导入数据
 - 提示: Excel 数据, 每个同学的 Y 数据都不同, 找到自己学号对应下的 Y 数据 (X 数据所有同学都一样)
 - 菜单操作 (Excel 和 Eviews):
 - a. Excel 找到数据。Excel 表格中仅保留自己需要的数据 (obs, wage, age, edu, dpt, sex)
 - b. Eviews 导入数据。File⇒Import⇒Import From File:d:/econometrics/data/Lab8-indian

把定性变量设置成虚拟变量体系

- 目标: 学会用一套虚拟变量体系来完整表达一个定性变量

- 思路：按照完备、互斥的法则设置虚拟变量；如果要设置有截距模型，应统筹、优先考虑基础组的虚拟变量设置。

- Eviews 操作：






- 1) 把定性变量教育水平 edu (m=4) 设置成虚拟变量体系。(具体操作见图??)

- a. 命令视窗 (Command) 依次输入命令 (建议分别命名为 edu_d1、edu_d2、edu_d3 和 edu_d4)

```
- series edu_d1=@recode(educ="illiteracy",1,0) '
- series edu_d2=@recode(educ="primary",1,0) '
- series edu_d3=@recode(educ="secondary",1,0)
- series edu_d4=@recode(educ="higher",1,0) '
```

- b. 运行命令：命令行中按 Enter 键

- c. 查看结果 (以组 group 的形式查看)：

- 按住键盘 Ctrl+ 依次点击 educ、edu_d1、edu_d2、edu_d3、edu_d4
- 点击鼠标右键 ⇒ Open ⇒ as Group

- 2) 把定性变量合同类型 dpt (m=2) 设置成虚拟变量体系。(具体操作见图??)

- a. 命令视窗 (Command) 依次输入命令 (建议分别命名为 dpt_d1 和 dpt_d2)

```
- series dpt_d1=@recode(dpt="temporary",1,0) '
- series dpt_d2=@recode(dpt="permanent",1,0)
```

- b. 运行命令：命令行中按 Enter 键

- c. 查看结果 (以组 group 的形式查看)：

- 按住键盘 Ctrl+ 依次点击 dpt、dpt_d1、dpt_d2
- 点击鼠标右键 ⇒ Open ⇒ as Group



- 3) 把定性变量性别 sex (m=2) 设置成虚拟变量体系。(具体操作见图??)

- a. 命令视窗 (Command) 依次输入命令 (建议分别命名为 sex_d1 和 sex_d2)





```
- series sex_d1=@recode(sex="female",1,0) '
- series sex_d2=@recode(sex="male",1,0)
```

- b. 运行命令：命令行中按 Enter 键

- c. 查看结果 (以组 group 的形式查看)：

- 按住键盘 Ctrl+ 依次点击 sex、sex_d1、sex_d2
- 点击鼠标右键 ⇒ Open ⇒ as Group

- 4) 说明 (Eviews 代码行的解读^[具体细节请参看 Eviews 在线学习文档，网址 <http://www.eviews.com/Learning/dummies.html>])：

- a. 代码 `series edu_d1=@recode(educ="primary",1,0)` 表示给创建一个序列 (Series) 对象 edu_d1，并对定性变量对象 educ 进行重新编码处理 (recode)，并把重新编码处理后的数值赋值给序列 (Series) 对象 edu_d1。
- b. 代码 `@recode(educ="primary",1,0)` 表示对定性变量对象 educ 进行重新编码处理。具体做法是，如果 educ 的取值为 primary，则相应赋值为 1，或者就相应赋值为 0。

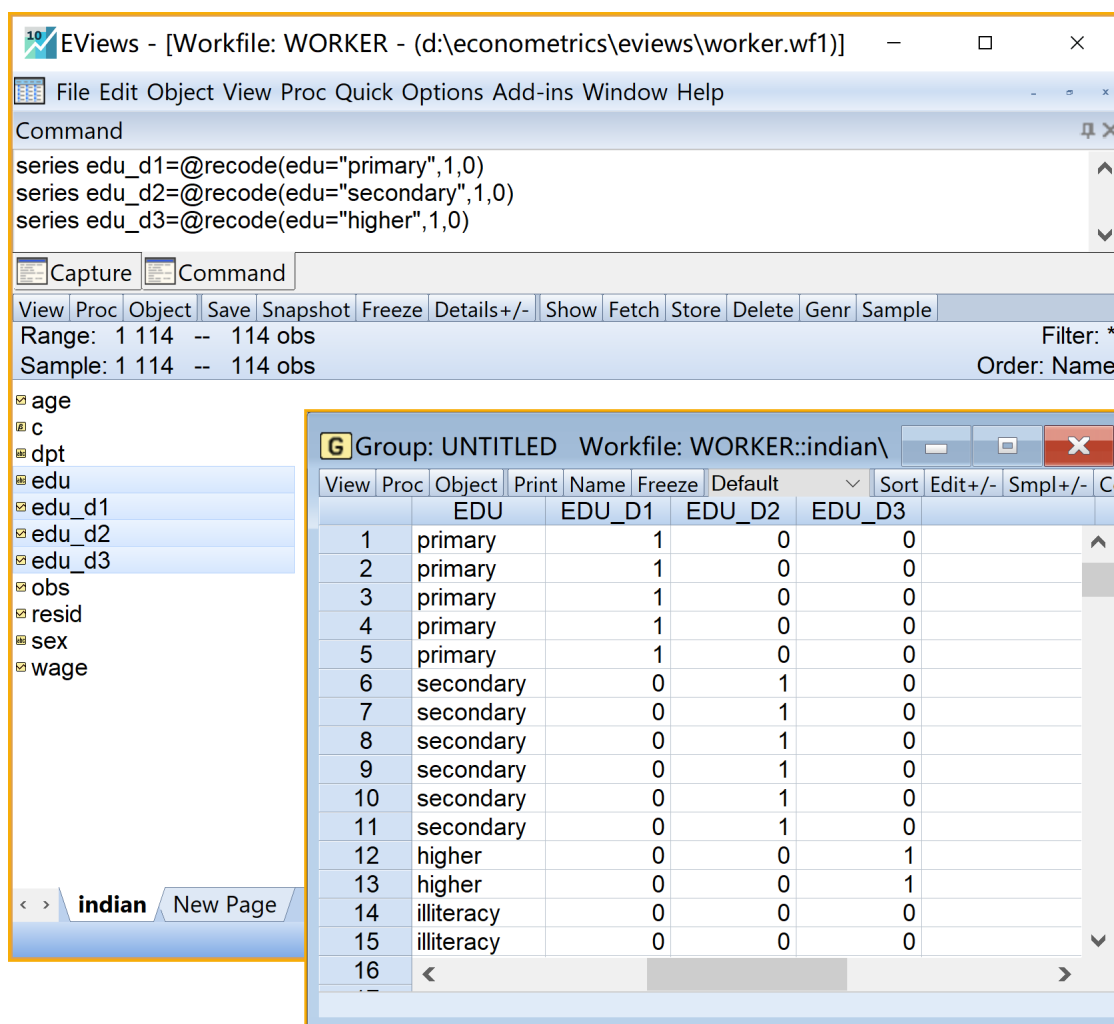


图 2: 定性变量 edu 用虚拟变量体系表达

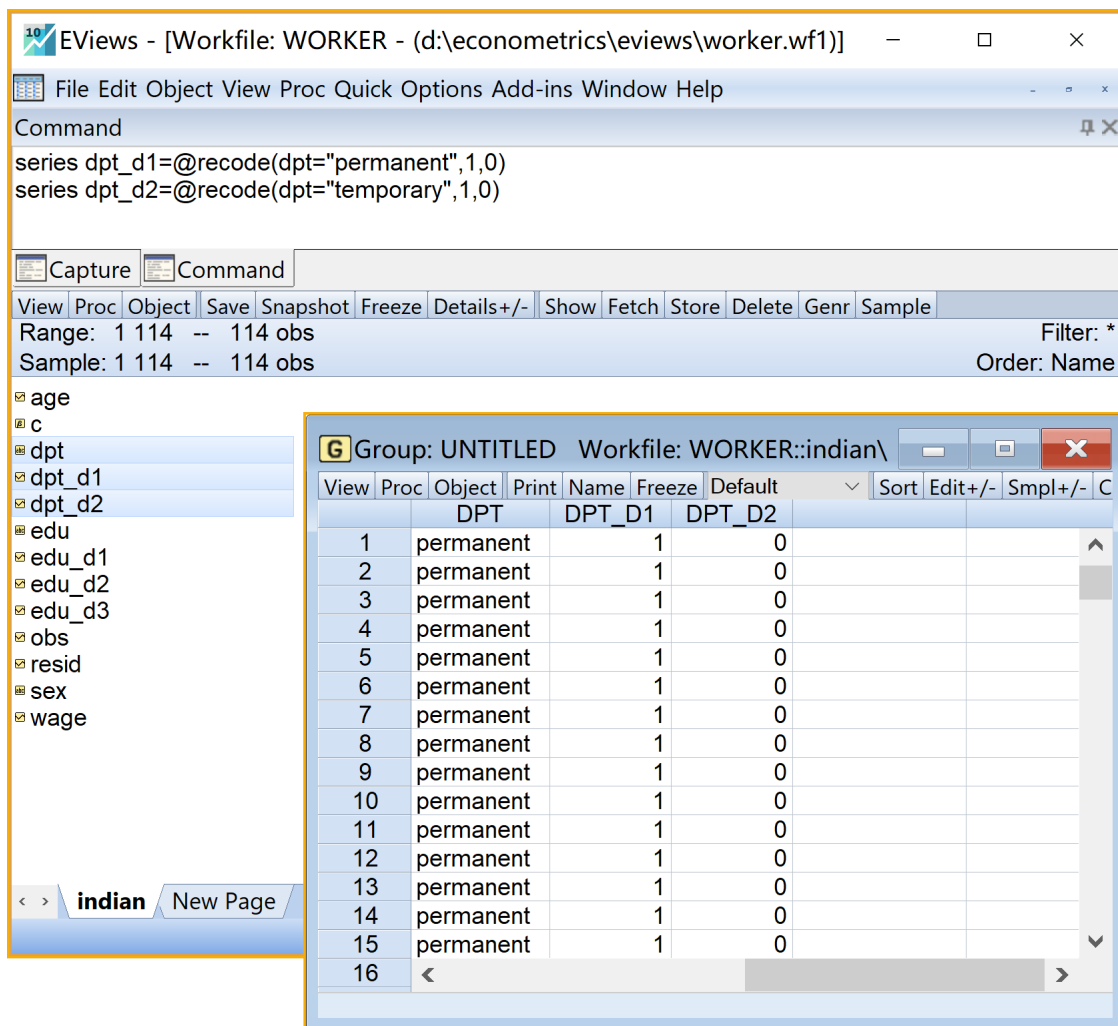


图 3: 定性变量 dpt 用虚拟变量体系表达

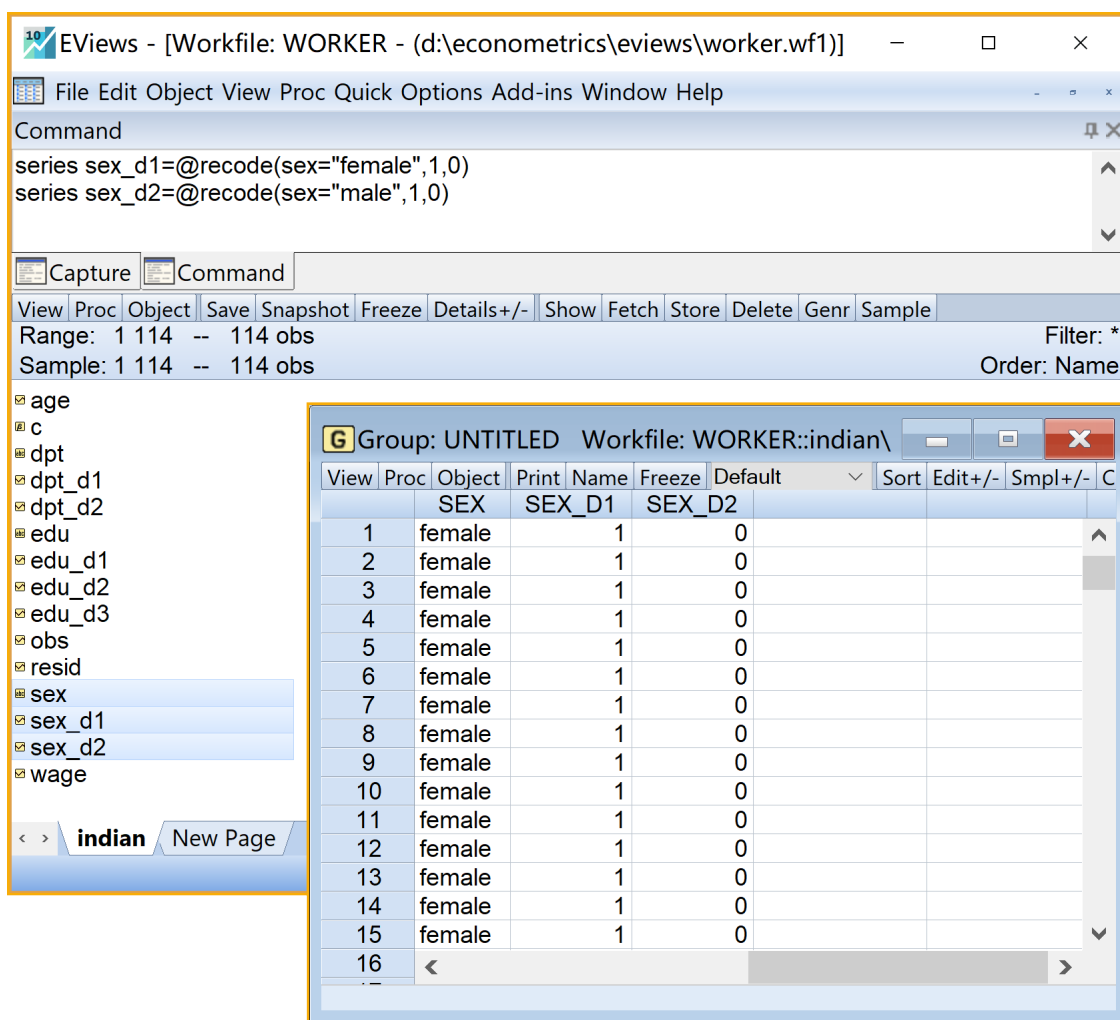


图 4: 定性变量 sex 用虚拟变量体系表达

表 3: 用虚拟变量系统完全表达定性变量 edu

	edu	edu_D1	edu_D2	edu_D3	edu_D4
1	primary	0	1	0	0
2	primary	0	1	0	0
6	secondary	0	0	1	0
7	secondary	0	0	1	0
12	higher	0	0	0	1
13	higher	0	0	0	1
14	illiteracy	1	0	0	0
15	illiteracy	1	0	0	0

表 4: 用虚拟变量系统分别完全表达定性变量 dpt 和 sex

	dpt	dpt_D1	dpt_D2		sex	sex_D1	sex_D2
1	permanent	0	1	1	female	1	0
2	permanent	0	1	2	female	1	0
3	permanent	0	1	3	female	1	0
112	temporary	1	0	112	male	0	1
113	temporary	1	0	113	male	0	1
114	temporary	1	0	114	male	0	1

操作解读

```
## Loading required package: psych
```

实际操作中，我们首先要对定性变量进行重新编码，设置成各自的虚拟变量体系。Eviews 中对定性变量重新编码为虚拟变量的代码函数为 `@recode()`。我们可以事先将一个定性变量完全地进行虚拟变量编码⁶。也就是说，如果一个定性变量有 m 个属性，我们可以直接设置 m 个虚拟变量。

此外，便于后续多个模型的分析甄选，我们还应该进一步统一设计虚拟变量的名称、命名的顺序等。例如，假设后续的备选模型中将基础组设定为 {文盲，临时工，女性}（也即 {illiteracy, temporary, female}）⁷。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_D1`（见表??）、`dpt_D1` 和 `sex_D1`（见表??）。

只含有虚拟变量的回归模型（考虑基础组的情形）

加法模型

⁶此时我们可以完全不用关心模型是否有截距（意味着是否有有对照比较的基础组）



⁷理论上，基础组如何选择并不会从根本上改变模型的实际经济学意义，只是一旦选定一个基础组，也就意味着确定了一个相互比较的“基础参照系”。

- 目标：把定性变量的虚拟变量以独立项的形式引入模型方程，解释回归报告
- 思路：确定基础组，设置总体回归模型（PRM），进行 OLS 估计，得到 Eviews 分析报告
- 理论提示：
 - 模型 1：只含有虚拟变量的、加法形式的经典回归模型见方程(??)
 - 模型 2：只含有虚拟变量的、加法形式的半对数回归模型见方程(??)


$$wage_i = \alpha_1 + \alpha_2 edu_{D2,i} + \alpha_3 edu_{D3,i} + \alpha_4 edu_{D4,i} + \beta_2 dpt_{D2,i} + \gamma_2 sex_{D2,i} + u_i \quad (5)$$

$$\ln(wage_i) = \alpha_1 + \alpha_2 edu_{D2,i} + \alpha_3 edu_{D3,i} + \alpha_4 edu_{D4,i} + \beta_2 dpt_{D2,i} + \gamma_2 sex_{D2,i} + u_i \quad (6)$$

(7)

- Eviews 操作 1（只含有虚拟变量的、加法形式的经典回归模型见方程(??)，菜单操作实现具体见图??）：
 - 1) 确定参照组为 [文盲 & 短期合同 & 女性]，则如下虚拟变量将不进入回归模型
 - a. ☒ edu_d1
 - b. ☒ dpt_d1
 - c. ☒ sex_d1
 - 2) 设置回归模型。进入引导设置 Equation Estimation \Rightarrow specification
 - a. Equation specification: 输入命令 wage c edu_d2 edu_d3 edu_d4 dpt_d2 sex_d2
 - b. Estimation settings:
 - Method: 下拉选择 LS - Least Squares (NLS and ARMA)
 - Sample: (默认设置)
 - c. 点击完成: OK
 - d. 命名保存方程对象 : (建议命名为 eq_only_plus)
 - e. 查看结果: 双击  eq_only_plus

具体 Eviews 报告见??:

- Eviews 操作 2（只含有虚拟变量的、加法形式的半对数回归模型见方程(??)，菜单操作实现具体见图??）：
 - 1) 确定参照组为 [文盲 & 短期合同 & 女性]，则如下虚拟变量将不进入回归模型
 - a. ☒ edu_d1
 - b. ☒ dpt_d1
 - c. ☒ sex_d1
 - 2) 设置回归模型。进入引导设置 Equation Estimation \Rightarrow specification
 - a. Equation specification: 输入命令 log(wage) c edu_d2 edu_d3 edu_d4 dpt_d2 sex_d2
 - b. Estimation settings:
 - Method: 下拉选择 LS - Least Squares (NLS and ARMA)
 - Sample: (默认设置)
 - c. 点击完成: OK
 - d. 命名保存方程对象 : (建议命名为 eq_only_plus_log)

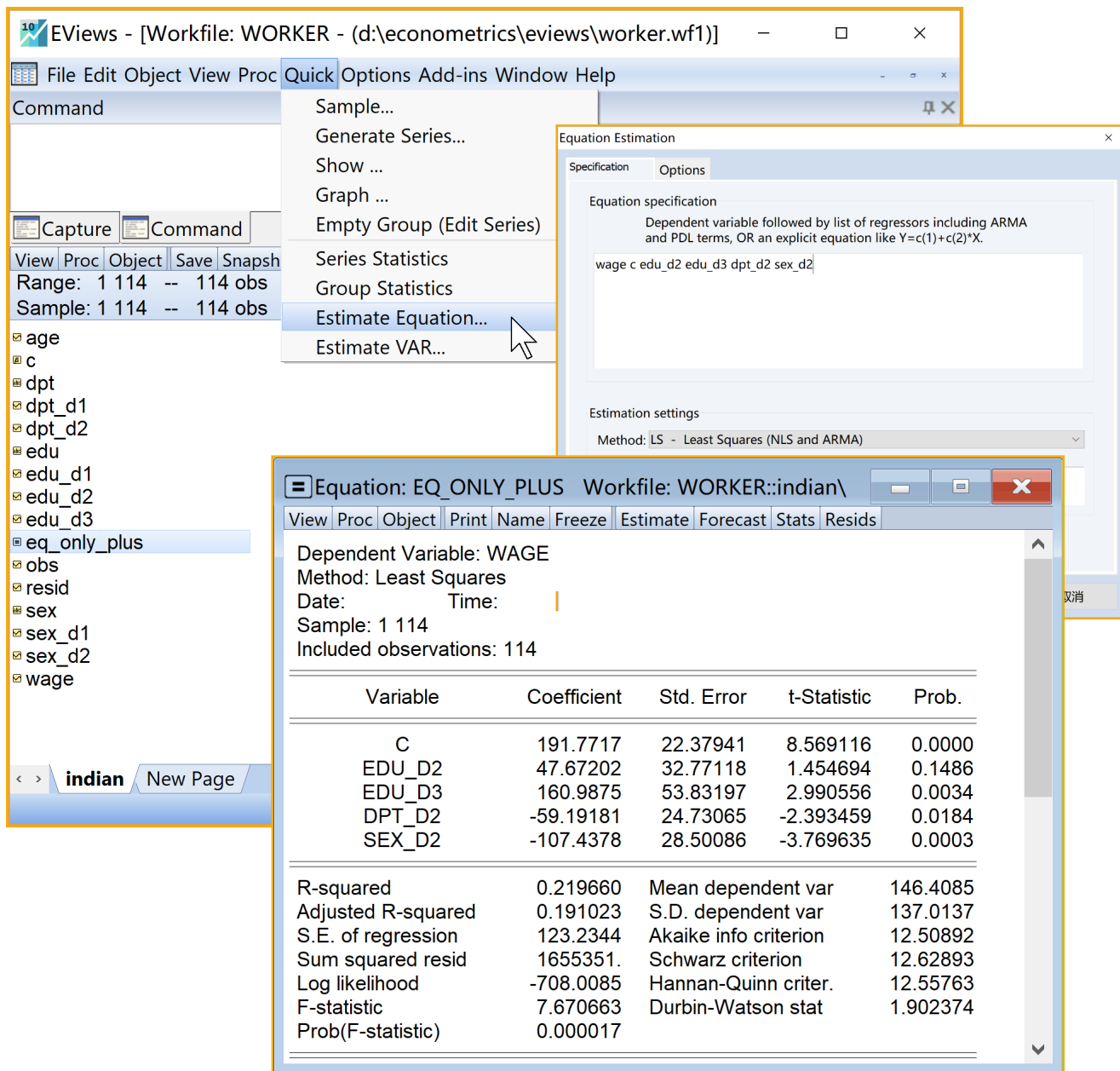


图 5: 只含虚拟变量的、加法形式的经典线性回归模型 Eviews 实现