homework 03: Endogeneity and IV method  
内生自变量问题及工具变量法

胡华平

2023-09-12

学生姓名：\_\_\_\_\_\_\_\_\_\_\_\_\_\_；学生学号：\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_；专业班级：\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**问：** 你将使用什么编程分析软件完成本次作业？

**答1：** 我选择的分析软件是：\_\_\_\_\_\_\_\_。

**答2：** 我使用的数据集文件名为：\_\_\_\_\_\_\_\_。（学号不在系统里的同学填写此空！）

**温馨提示**：建议大家优先选择R、python、stata。当然，也可以使用Matlab或EViews的编程功能。

# 作业提交

**作业发布时间**：2023-09-12（周二）24:00:00

**提交截止时间**：2023-09-24（周日）24:00:00

**作业提交材料**：

（1）根据作业要求，完成Office Word电子文档一份（注意不能是wps文档），提交前请将文件命名为下述格式：lab03\_word\_张三\_2019000001.docx。

（2）根据作业要求，完成相关编程分析操作，保存并提交1份**原始代码文件**（其中10位数字为学生的学号）：

* 如果使用R软件，请将编程代码文件保存并命名为：lab03\_code\_Rscript\_2019000001.r；或者lab03\_code\_Rmarkdown\_2019000001.Rmd。
* 如果使用Python软件，请将编程代码文件保存并命名为：lab03\_code\_Python\_2019000001.py。
* 如果使用Stata软件，请将编程代码文件保存并命名为：lab03\_code\_Stata\_2019000001.do。
* 如果使用Matlab软件，请将编程代码文件保存并命名为：lab03\_code\_Matlab\_2019000001.m。
* 如果使用EViews软件，请将编程代码文件保存并命名为：lab03\_code\_EViews\_2019000001.prg。

**作业提交方式**：

* 按上述要求命名各个文件，然后将全部作业提交材料压缩为zip格式，并命名为：lab03-专业年级-姓名-学号.zip（如：lab03-2023应经-张帅帅-2023120208.zip）
* 发送上述zip文件到电子邮箱：huhuaping01@qq.com。请填写“邮件主题”为：lab03-专业年级-姓名-学号（如：lab03-2023应经-张帅帅-2023120208）。
* 点击发送邮件！你将收到一份邮件已收到的“自动回复”！

# 作业案例

Warning: Using `all\_of()` outside of a selecting function was deprecated in tidyselect  
1.2.0.  
ℹ See details at  
 <https://tidyselect.r-lib.org/reference/faq-selection-context.html>  
This warning is displayed once every 8 hours.  
Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
generated.

## 变量说明

为了研究居民的工资收入（wage，或对数化的工资lwage）是如何决定的，我们考虑如下所示的一些变量（具体定义见表1）。

Table 1: 变量定义及说明

| 变量\_代码 | 变量\_中文 | 定义和取值 |
| --- | --- | --- |
| obs | 序号 | 序号 |
| lwage | 对数化工资 | 定量变量：取对数后的工资 |
| edu | 受教育年数 | 定量变量：受教育年数（年） |
| exp | 工作年数 | 定量变量：工作年数(年) |
| exp2 | 工作年数平方 | 定量变量：工作年数平方/100 |
| black | 是否黑人 | 虚拟变量：1=黑人；0=其他 |
| south | 是否南方地区 | 虚拟变量：1=南方地区；0=其他 |
| urban | 是否居住市区 | 虚拟变量：1=居住市区；0=其他 |
| college | 附近是否有大学 | 虚拟变量：1=附近有大学；0=其他 |
| public | 附近是否有公立大学 | 虚拟变量：1=附近有公立大学；0=其他 |
| private | 附近是否有私立大学 | 虚拟变量：1=附近有私立大学；0=其他 |
| age | 年龄 | 定量变量：年龄(年) |
| age2 | 年龄平方 | 定量变量：年龄(年)的平方/100 |
| momedu | 母亲受教育年数 | 定量变量：母亲受教育年数（年） |
| dadedu | 父亲受教育年数 | 定量变量：父亲受教育年数（年） |

## 数据说明

**重要说明**：

* 每一个同学的数据集都不同。其中acdemic表示**学硕学生**；major表示**专硕学生**。
* 数据文件利用学号进行命名，例如card1995\_major\_2023056341.xlsx。
* 大家根据自己的学号，使用相应的数据集。
* 考虑到部分同学的学号没有在系统里，我虚拟了15个学号（从2022056001-2022056015）。这部分同学可以任意使用其中的一份数据集开始自己的作业。

同时，我们给每一个同学**分别**准备了一份关于工资收入的配套数据集（数据文件名类似如card1995\_major\_major056341.xlsx），该数据为截面数据（样本量n=3000），下面的图1给出了一个示例。

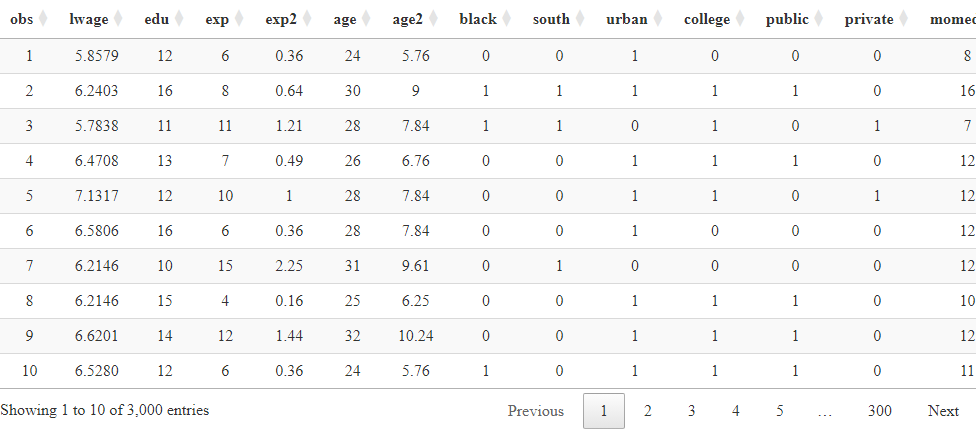


Figure 1: 样本数据集示例 n=3000

# 作业任务

## 任务1：OLS回归分析

**任务内容：** OLS方法作为一种经典估计方法，任何一款编程软件都能在数据集和指定模型下，快速实现OLS参数估计，并报告出分析结果。请你使用自己的编程软件，根据下列多元线性回归模型（见式(1)），利用特定**函数命令代码**进行快速回归分析，并完成如下相关任务。

### **任务1-1：** OLS命令应用

给定数据集为dt\_card，以及回归模型为mod\_ols。根据总体回归模型（见式(1)），请在下面写出你编程软件下快速OLS回归分析的代码命令（一行命令）：

**答：** 代码命令为：

**温馨提示**：例如在R编程软件下，OLS回归分析的函数命令为lm\_est <- lm(formula = mod\_ols, data = dt\_card)。

### **任务1-2：** OLS估计结果的整理

（1）请你将上述OLS分析报告的估计结果整理成表格形式（具体格式见下面的**温馨提示**，保留4位小数）。

**答：** OLS分析报告的表格化整理结果为：

**温馨提示**：例如，线性模型，OLS方法估计结果可以整理为如下表格形式（见表2）：

Table 2: OLS估计结果表格化整理演示

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | 4.4685 | 0.0687 | 65.0538 | 0 |
| edu | 0.0932 | 0.0036 | 26.0237 | 0 |
| exp | 0.0898 | 0.0071 | 12.7106 | 0 |
| exp2 | -0.2486 | 0.0338 | -7.3609 | 0 |

（2） 使用公式编辑器（或Mathtype软件），将上述分析报告，手动整理成**样本回归模型**（SRM）的简要报告（四行报告，包括第1行样本回归方程、第2行对应的系数标准误、第3行对应的样本t统计量，以及第4行F检验值、p值、拟合优度等。具体形式见课件及后面提示），将结果填写在下面空白处。

**答：**OLS分析报告的多行式整理结果为：

**温馨提示**：例如，线性模型，OLS方法估计得到的样本回归模型可以整理如下：

## 任务2：一个内生变量且等量工具变量情形下的2SLS分析（edu内生，工具变量为college）

**任务内容：**假如我们认为前述OLS模型（见式(1)）中受教育年数edu为内生自变量，并且认为其他自变量都是外生的。同时，我们选择把附近是否有大学college作为内生自变量edu的一个备选工具变量。请你进行两阶段最小二乘回归（2SLS）分析（具体两阶段模型见下面的式(2)和式(3)），并回答相关问题：

### **任务2-1：** 手动分步2SLS法

对于上述两阶段关联模型（见式(2)和式(3)），如果你打算自己手动分两步进行工具变量估计（手动两阶段最小二乘法），也即：第一步先用OLS方法估计式(2)，得到；第二步，把之前估计得到的替换为式(3)中的，然后再进行第二阶段的OLS回归。

**（1）**请你写出第一阶段分析的样本回归函数（SRF）的理论表达式，同时将它的OLS估计结果整理成表格形式（具体格式见前面的**温馨提示**，保留4位小数）。

**答：** 第一阶段分析的样本回归函数（SRF）的理论表达式为：

第一阶段OLS估计进行表格化整理后的结果为：

**（2）**请你写出第2阶段分析的样本回归模型（SRM）的理论表达式，同时将它的OLS估计结果整理成表格形式（具体格式见前面的**温馨提示**，保留4位小数）。

**答：** 第二阶段分析的样本回归模型（SRM）的理论表达式为：

第二阶段OLS估计进行表格化整理后的结果为：

### **任务2-2：** 快速2SLS法

对于上述两阶段关联模型（见式(2)和式(3)），我们也可以使用编程软件内置的命令进行快速2SLS估计，从而同时得到两个回归方程的最终估计结果。

（1）根据上述两阶段关联模型（见式(2)和式(3)），请在下面写出你编程软件下快速2SLS回归分析的代码命令：

**答：** 相关代码命令为：

**温馨提示**：例如，两阶段关联模型和，快速2SLS估计的R命令代码如下：

# load package  
library(systemfit)  
# set two models  
eq\_1 <- edu ~ exp + exp2 + college  
eq\_2 <- lwage ~ edu + exp + exp2   
sys <- list(eq1 = eq\_1, eq2 = eq\_2)  
  
# specify the instruments  
instr <- ~ exp + exp2 + college  
  
# fit models   
fit.sys <- systemfit(  
 formula = sys,   
 inst = instr,   
 method ="2SLS",   
 data = tbl\_reg)

**（2）**请你将上述快速2SLS法估计结果整理成表格形式（具体格式见下面的**温馨提示**，保留4位小数）。

**答：** 快速2SLS法估计进行表格化整理后的结果为：

**温馨提示**：例如，两阶段关联模型和，快速2SLS估计结果可以整理为如下表格形式（见表3）：

Table 3: 快速2SLS估计结果表格化整理演示

| model | term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- | --- |
| eq1 | (Intercept) | 16.5734 | 0.1702 | 97.3932 | 0.0000 |
| eq1 | exp | -0.4225 | 0.0348 | -12.1353 | 0.0000 |
| eq1 | exp2 | 0.0235 | 0.1704 | 0.1379 | 0.8903 |
| eq1 | college | 0.6002 | 0.0788 | 7.6168 | 0.0000 |
| eq2 | (Intercept) | 1.6540 | 0.5802 | 2.8509 | 0.0044 |
| eq2 | edu | 0.2587 | 0.0340 | 7.6012 | 0.0000 |
| eq2 | exp | 0.1597 | 0.0170 | 9.4094 | 0.0000 |
| eq2 | exp2 | -0.2488 | 0.0442 | -5.6306 | 0.0000 |

### **任务2-3：** 两种方法的结果比较

通过**手动分步2SLS法**（见前面任务2-1）和**快速2SLS法**（见前面任务2-2），我们分别能得到第二阶段的估计结果（也即式(3)）。请你简要说明两种方法得到的结果有什么异同之处？如果有不同之处，主要的原因是什么？

**答：** 两种方法的结果主要异同之处包括：

## 任务3：两个内生变量且等量工具变量情形下的2SLS分析（edu、exp和exp2内生，工具变量为college、age和age2）

**任务内容：**假如我们认为前述OLS模型（见式(1)）存在内生性问题，其中受教育年数edu、工作年数及其平方exp和exp2为内生自变量，并且认为其他自变量都是外生的。同时，我们选择把附近是否有大学college作为内生自变量edu的一个备选工具变量，把年龄及其平方age和age2分别作为exp和exp2的工具变量。请你进行两阶段最小二乘回归（2SLS）分析（具体两阶段模型见下面的式(4)、式(5)、式(6)，以及式(7)），并回答相关问题：

### **任务3-1：** 内生性的来源及其可选工具变量

对于上述两阶段关联模型（见上面的式(4)、式(5)、式(6)，以及式(7)），请你简要解释为什么工作年数及其平方exp和exp2会产生内生性问题？以及为什么年龄及其平方age和age2可以分别作为它们的可选工具变量？

**答：**

### **任务3-2：** 快速2SLS法

对于上述两阶段关联模型（见上面的式(4)、式(5)、式(6)，以及式(7)），我们也可以使用编程软件内置的命令进行快速2SLS估计，从而同时得到两个回归方程的最终估计结果。

（1）根据上述两阶段关联模型（见上面的式(4)、式(5)、式(6)，以及式(7)），请在下面写出你编程软件下快速2SLS回归分析的代码命令：

**答：**相关代码命令为：

**（2）**请你将上述快速2SLS法估计结果整理成表格形式（具体格式同前面的**温馨提示**，保留4位小数）。

**答：** 快速2SLS法估计进行表格化整理后的结果为：

### **任务3-3：** 内生变量之间的关系

根据常识以及前面数据集和变量说明，我们易知：对于大多数人来说，他的教育年数edu和他工作年数exp本身存在一个等式关系：，也即工作年数等于其年龄减去受教育年数（因为大多数人毕业后就会选择去工作）。基于这一事实，请你回答如下问题：

（1）结合**任务3-2**的估计结果，请你分别抄写出**第1阶段回归**中，教育年数edu回归模型（见式(4)）和工作年数exp回归模型（见式(5)）的相关结果，请你填写完成下表（见表4）（要求：保留4位小数）。

**答：** 完成下表下划线填空即可。其中括号前为估计系数，括号内为估计标准误差，表格补充填写如下（要求：保留4位小数）：

Table 4: 请填写回归系数及标准误差（括号内）

| 解释变量 | 被解释变量edu | 被解释变量exp |
| --- | --- | --- |
| black | \_\_\_\_(\_\_\_\_) | \_\_\_\_(\_\_\_\_) |
| south | \_\_\_\_(\_\_\_\_) | \_\_\_\_(\_\_\_\_) |
| urban | \_\_\_\_(\_\_\_\_) | \_\_\_\_(\_\_\_\_) |
| age | \_\_\_\_(\_\_\_\_) | \_\_\_\_(\_\_\_\_) |
| age2 | \_\_\_\_(\_\_\_\_) | \_\_\_\_(\_\_\_\_) |
| (Intercept) | \_\_\_\_(\_\_\_\_) | \_\_\_\_(\_\_\_\_) |

（2）根据上述填写完整的表格结果（见表4），对于有关变量（black、south、urban、college）的估计结果，你有没有发现什么有趣的结论？为什么会出现这样的结果？对于前面讨论的等式关系（），你认为这是否会导致**第1阶段回归**中，教育年数edu回归模型（见式(4)）和工作年数exp回归模型（见式(5)）二者之一，必定有一个是多余的？请简要说明你的理由！

**答：**

## 任务4：一个内生变量且富余工具变量情形下的2SLS分析（edu内生，工具变量为public和private）

**任务内容：**这里，假如我们认为前述OLS模型（见式(1)）存在内生性问题，其中受教育年数edu为内生自变量，并且认为其他自变量都是外生的。同时，我们选择把附近是否有公立大学public和附近是否有私立大学private作为内生自变量edu的2个备选工具变量。请你进行两阶段最小二乘回归（2SLS）分析（具体两阶段模型见下面的式(8)和式(9)），并回答相关问题：

### **任务4-1：** 多个可选工具变量的关系

对于上述两阶段关联模型（见式(8)和式(9)），请你简要解释为什么我们可以把附近是否有公立大学public和附近是否有私立大学private作为内生自变量edu的2个备选工具变量？进一步地，我们能不能将3个外生变量college、public和private同时作为内生自变量edu的备选工具变量，并将三者同时放到第一阶段模型（见式(8)）中去？请简要说明理由！

**答：**

### **任务4-2：** 快速2SLS法

对于上述两阶段关联模型（见式(8)和式(9)），我们也可以使用编程软件内置的命令进行快速2SLS估计，从而同时得到两个回归方程的最终估计结果。

（1）根据上述两阶段关联模型（见式(8)和式(9)），请在下面写出你编程软件下快速2SLS回归分析的代码命令：

**答：**相关代码命令为：

**（2）**请你将上述快速2SLS法估计结果整理成表格形式（具体格式同前面的**温馨提示**，保留4位小数）。

**答：** 快速2SLS法估计进行表格化整理后的结果为：

### **任务4-3：** 不同工具变量选择下的模型比较

自此，仅就内生自变量edu而言，我们重点考虑了被调查附近是否有大学（college）以及大学类型（public和private），作为其备选工具变量进行分别进行了2SLS建模分析，并得到了估计结果（分别见**任务2-2**和**任务4-2**）。

（1）请你对二者的**第1阶段**回归结果进行分析。其中，edu对college的第一阶段回归模型估计结果（具体见式(2)以及见**任务2-2**的表格结果）；edu对public和private的第一阶段回归模型估计结果（具体见式(8)以及见**任务4-2**的表格结果）。你认为被调查者受教育年数edu是如何受到他出生地附近是否有大学college以及大学是哪种类型（public和private）影响的？（要求：对比估计量，并做出说明！）

**答：**

（2）进一步地，请你对二者的**第2阶段**回归结果进行分析。其中，edu对college的第一阶段回归模型估计结果（具体见式(3)以及见**任务2-2**的表格结果）；edu对public和private的第一阶段回归模型估计结果（具体见式(9)以及见**任务4-2**的表格结果）。你认为被调查者出生地附近是否有大学college以及大学是哪种类型（public和private），又是如何进一步影响到他的工资收入lwage的？你能得到怎样的结论？（要求：对比估计量，并做出说明！）

**答：**

## 任务5：两个内生变量且富余工具变量情形下的2SLS分析（edu、exp和exp2内生，工具变量为public和private、以及age和age2）

**任务内容：**这里，假如我们认为前述OLS模型（见式(1)）存在内生性问题，其中受教育年数edu、工作年数及其平方exp和exp2为内生自变量，并且认为其他自变量都是外生的。同时，我们选择把附近是否有公立大学public和附近是否有私立大学private作为内生自变量edu的2个备选工具变量；把年龄及其平方age和age2分别作为exp和exp2的工具变量。请你进行两阶段最小二乘回归（2SLS）分析（具体两阶段模型见下面的式(10)、式(11)、式(12)和式(13)），并回答相关问题。

### **任务5-1：** 工具变量选择及数量问题

对于上述两阶段关联模型（见上面的式(10)、式(11)、式(12)和式(13)），请你回答哪些是内生变量？哪些是外生变量？进一步地，对于我们的实证分析，工具变量是多多益善么？（要求：写出变量名称，做出简要说明）

**答：**

### **任务5-2：** 快速2SLS法

对于上述两阶段关联模型（见上面的式(10)、式(11)、式(12)和式(13)），我们也可以使用编程软件内置的命令进行快速2SLS估计，从而同时得到两个回归方程的最终估计结果。

（1）根据上述两阶段关联模型（见上面的式(10)、式(11)、式(12)和式(13)），请在下面写出你编程软件下快速2SLS回归分析的代码命令：

**答：**相关代码命令为：

**（2）**请你将上述快速2SLS法估计结果整理成表格形式（具体格式同前面的**温馨提示**，保留4位小数）。

**答：** 快速2SLS法估计进行表格化整理后的结果为：

### **任务5-3：** 工作年数是否显著影响工资收入？

对于**第2阶段**回归模型（具体见式(13)及其估计结果（见**任务5-2**）。请你指出，变量exp和exp2系数估计的t检验结果是否显著（给定）？那么，工作年数是否显著影响工资收入？这与我们的理论或经验相一致么？（要求：对比估计量及显著性检验，并做出说明！）

**答：**

## 任务6：模型综合比较

**任务内容：**至此，我们最终总共得到了五种工资影响因素的分析模型及结果。具体为：

* 模型1：OLS回归，完全不考虑是否有内生性问题。工资影响模型见式(1)，其OLS估计结果见**任务1-2**。
* 模型2：一个内生变量且等量工具变量2SLS分析。其中考虑了edu内生，并使用工具变量college。工资影响模型见式(3)，其第二阶段估计结果见**任务2-3**。
* 模型3：2个内生变量且等量工具变量的2SLS分析。其中考虑edu、exp和exp2内生，工具变量为college、age和age2。工资影响模型见式(7)，其第二阶段估计结果见**任务3-3**。
* 模型4：一个内生变量且富余工具变量的2SLS分析。其中考虑edu内生，工具变量为public和private。工资影响模型见式(9)，其第二阶段估计结果见**任务4-2**。
* 模型5：两个内生变量且富余工具变量的2SLS分析。考虑edu、exp和exp2内生，工具变量为public和private、以及age和age2。工资影响模型见式(13)，其第二阶段估计结果见**任务5-2**。

### **任务6-1：** 整理不同工资决定模型的结果

根据上述五个工资决定模型，以及前面已经完成的估计结果，请你将全部模型结果整理为一张表（具体表格样式见下面的温馨提示）。（要求：仅列出估计系数及对应的标准误）。

**答：** 全部模型结果可以整理为如下的一张表（保留4位小数，表格样式见下面的温馨提示）：

**温馨提示**：以下为整理的结果（见图2），每一列分别与上述五个工资模型（因变量为lwage）的情景相对应。其中，括号前为估计系数，括号内为估计标准误。此表结果和样式仅供参考，具体数值以自己的计算为准。

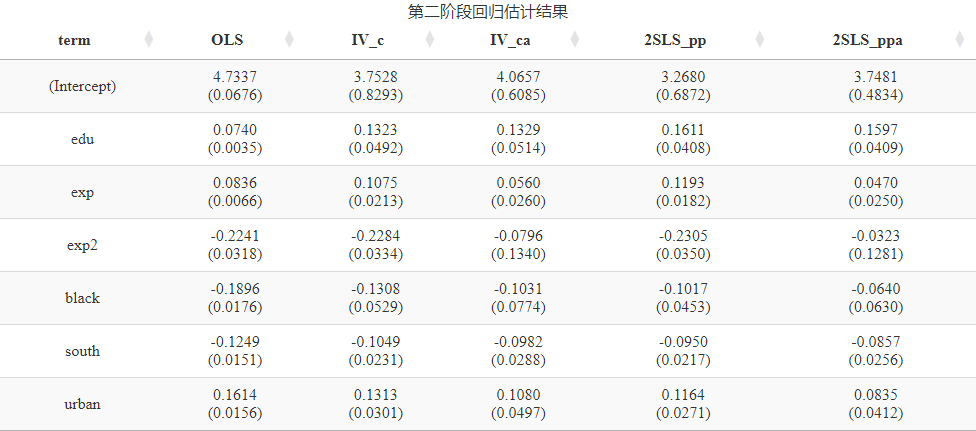


Figure 2: 结果整理的表格样式示例（仅供参考）

### **任务6-2：** 对比工资决定模型的估计差异

根据上述工资决定模型估计结果的整理，请你依次回答：

（1）请简要说明教育年数edu对工资lwage的影响，在五个模型中有什么表现及不同？

**答：**

（2）请简要说明是否黑人black对工资lwage的影响，在五个模型中有什么表现及不同？肤色（黑人）对工资的影响真的如大多数模型结果所估计的那么大么？五个模型对这一系数值的估计是稳定和一致的么？

**答：**