homework week02: Multiple Linear Regression Programming  
多元线性回归矩阵方法及其编程

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：** 我选择的分析软件是：\_\_\_\_\_\_\_\_。

**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**：

**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：lab02\_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：lab02\_code\_Rscript\_2019000001.r；或者lab02\_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：lab02\_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：lab02\_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：lab02\_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：lab02\_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：lab2-专业年级-姓名-学号.zip（如：lab2-2023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：lab2-专业年级-姓名-学号（如：lab2-2023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

## 变量和数据

为了研究居民的工资收入（wage，或对数化的工资lgwg）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）。同时，我们给出了一份关于工资收入的配套数据集（cps.xlsx），该数据为截面数据（样本量n=46943），下面的图1做出了部分展示。

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 |
| wage | 工资 | 定量变量：美元 |
| lgwg | 对数化工资 | 取对数后的工资 |
| edu | 受教育年数 | 定量变量：受教育年数（年） |
| exper | 工作年数 | 定量变量：工作年数(年) |
| exper2 | 工作年数平方 | 定量变量：工作年数平方/100 |
| female | 是否女性 | 虚拟变量：1=女性；0=男性 |
| union | 工会成员状态 | 定性变量：female=女性工会成员；male=男性工会成员；other=其他 |
| married | 已婚状态 | 定性变量：female=已婚女性；male=已婚男性；other=其他 |
| formMarried | 过往婚史状态 | 定性变量：female=再婚女性；male=再婚男性；other=其他 |
| hispanic | 是否拉美裔 | 虚拟变量：1=是；0=否 |
| race | 种族类别 | 定性变量：black=黑人；AI=印第安人；asian=亚裔；mixed=混血；other=其他 |

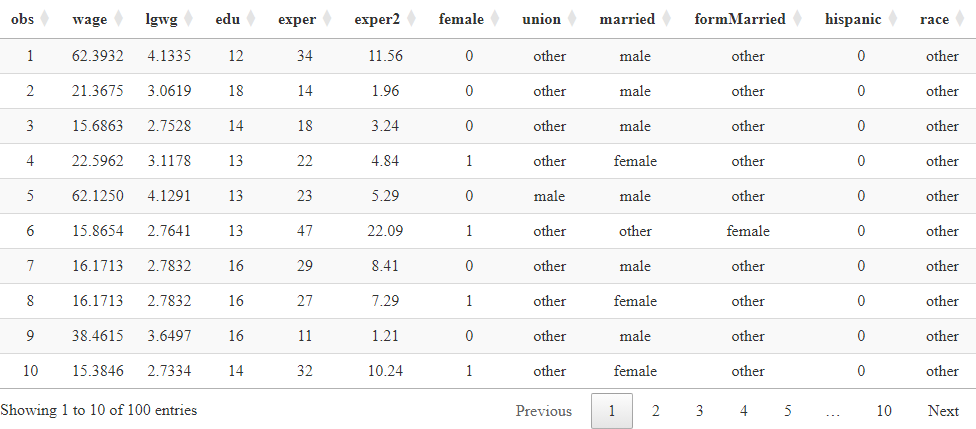


Figure 1: 样本数据集 n=46943

## 模型说明

本次作业我们将使用配套数据，进行如下的多元线性回归建模（虚拟变量模型）。

# 作业任务

## 任务1：处理定性变量

**任务内容：** 根据多元线性回归模型（见式(1)），请你对配套数据集中的**定性变量**，分别设置相应的虚拟变量体系。

**答：**此题无需作答，利用编程软件完成相关操作即可（变量处理）。

**温馨提示1**：如何把定性变量转换为虚拟变量？可以参看我的指导，网页链接 <https://book.huhuaping.com/dummy-variable.html#%E5%A6%82%E4%BD%95%E6%8A%8A%E5%AE%9A%E6%80%A7%E5%8F%98%E9%87%8F%E8%BD%AC%E6%8D%A2%E4%B8%BA%E8%99%9A%E6%8B%9F%E5%8F%98%E9%87%8F>

**温馨提示2**：例如，已婚状态变量married为定性变量，它具有三种取值属性（female=已婚女性；male=已婚男性；other=其他），因此可以将其处理为如下三个虚拟变量married\_male、married\_female、married\_other。考虑到我们的目标模型（见式(1)），实际上我们回归分析中，仅仅使用前两个虚拟变量（married\_male和married\_female）。

## 任务2：OLS回归估计（矩阵理论公式）

**任务内容：**我们可以使用相关矩阵理论公式，完成多元回归分析中的相关计算过程。请你依次完成如下提问：

（1）根据总体回归模型（见式(1)），请你填写补充如下的样本回归模型的矩阵表达式。

**答：**（要求写出矩阵的维度即可）

（2）根据总体回归模型（见式(1)），请写出OLS估计方法下，回归系数的矩阵理论计算公式：

**答：**

（3）根据总体回归模型（见式(1)），请写出OLS估计方法下，回归误差方差和回归误差标准差的矩阵理论公式：

**答：** 补充完整下列理论公式（注意**下横线**）：

（4）根据总体回归模型（见式(1)），请写出OLS估计方法下，请写出在符合**N-CLRM**假设情况下，回归系数的方差协方差矩阵的理论表达式。

**答：** 它的理论表达式为（请你补充完整）：

## 任务3：OLS回归估计（快速法）

**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，利用特定**函数命令代码**进行快速回归分析，并完成如下提问：

**任务3-1：** 给定数据集为dt\_cps，以及回归模型为mod\_dummy。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\_dummy, data = dt\_cps)。

**任务3-2：** 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：**

**温馨提示**：例如，线性模型，OLS方法估计得到的样本回归模型如下：

## 任务4：OLS回归估计（矩阵法）

**任务内容：** 当然，我们可以直接使用代数或矩阵计算公式（见**任务2**），利用编程命令一步步计算OLS的每一个细节过程，从而有助于更好地理解OLS分析的全过程。下面请你直接使用编程代码完成如下任务：

**任务4-1：** 根据总体回归模型（见式(1)），请你根据理论矩阵公式，使用**编程代码**计算得到回归系数，并将主要代码写在下面：

**答：** 回归系数矩阵计算的主要代码如下：

**温馨提示**：例如在R编程下，代码如下：

# step 1: create matrix y  
y <- as.matrix(dt\_cps$lgwg)  
  
# step 2: create matrix x  
## --- all variables in matrix X  
vars\_tar <- c(  
 "c", # attention here! constants with all ones !  
 "edu", "exper", "exper2",  
 "female",   
 "union\_female", "union\_male",  
 "married\_female","married\_male",  
 "formMarried\_female","formMarried\_male",  
 "hispanic",   
 "race\_black", "race\_AI","race\_asian", "race\_mixed")  
## --- now get the matrix X  
x <- dt\_cps %>%  
 select(all\_of(vars\_tar)) %>%  
 as.matrix()  
  
# step 3: finally calculate the beta\_hat matrix  
b <- solve(t(x)%\*%x)%\*%(t(x)%\*%y)

**任务4-2：** 根据OLS方法下回归系数的样本方差协方差矩阵的理论公式（见前面的**任务2-(4)**），请你使用**编程代码**计算得到这个矩阵的估计结果。然后你需要：

（1）把这个矩阵的对角线元素值提取出来，并开根号，形成一个列向量（实际上就是回归系数的估计标准误差），并填写在下列空白处。

（2）将主要代码写在下面空白处。

**答（1）：**回归系数的估计标准误差行向量为（填写空白处，保留4位小数）：

**答（2）：** 主要计算过程的编程代码为：

**温馨提示**：假设已经得到样本协方差矩阵为S2，提取矩阵对角线元素并开根号的操作（假设向量对象保存为S\_b）：例如在R编程语言下，代码命令为S\_b <- sqrt(diag(S2))；在EViews编程语言下的代码命令为matrix S\_b=@sqrt(@getmaindiagonal(S2))。

## 任务5：异方差一致性标准误差（选作题，有一定难度）

**任务内容**：对于总体回归模型（见式(1)），考虑到我们的数据集为截面数据，OLS估计结果很可能会出现**异方差问题**。也即意味着系数估计的样本标准误会不准确，从而无法达到统计推断（如t检验）的目的。请你采用怀特异方差一致性标准误方法，对样本标准误进行修正，得到修正后的结果。分别回答如下问题：

**任务5-1**：请你使用编程软件的**快速方法**（现成的函数命令）进行**怀特异方差一致性标准误矫正**，计算矫正后的方差协方差**矩阵**，提取矩阵对角线**元素**，然后对**向量**进行开根号计算，并将矫正后的系数标准误差**向量**保存为S1\_quick。**要求**：在下面空白处，写出矫正结果值（请转置为行向量，保留4位小数），并写出主要命令代码。

**答5-1：**

（1）怀特法矫正后回归系数的估计标准误差的行向量为（填写空白处，保留4位小数）：

（2）主要计算过程的编程代码为：

**任务5-2**：请你使用编程软件并利用**矩阵方法**（自己编写命令）进行**怀特异方差一致性标准误矫正**，计算矫正后的方差协方差**矩阵**，提取矩阵对角线**元素**，然后对**向量**进行开根号计算，并将矫正后的系数标准误差**向量**保存为S1。**要求**：比较S1与S1\_quick的计算结果是否一致（两种方法结果应该相同）！在下面空白处，写出主要命令代码。

**答5-2：** 二者计算结果【是/否】一致。主要计算过程的编程代码为：

**温馨提示1**：怀特异方差稳健协方差矩阵（Eicker-White or White covariance matrix estimator）一般记为HC0。其矩阵理论计算公式为：

**温馨提示2**：在R编程软件下**快速方法**的主要命令代码如下：

# step 1: Load necessary packages   
library("lmtest")  
library("sandwich")  
  
# step 2: OLS estimation  
lm\_est <- lm(formula = mod\_dummy, data = dt\_cps)  
  
# step 3: White robust covariance estimates for the coefficients, we get the covariance matrix  
vcov\_white <- vcovHC(lm\_est, type = "HC0")  
  
# step 4: quick show the standard error (S1)  
coeftest(lm\_est, vcov = vcov\_white )

**温馨提示3**：在R编程软件下**矩阵方法**的主要命令代码如下：

#=== continue after we finished estimate beta  
  
# step 1: get residual  
e <- y-x%\*%b  
leverage <- rowSums(x\*(x%\*%solve(t(x)%\*%x)))  
  
# step 2: get regression variance   
n <- nrow(y)  
k <- ncol(x)  
a <- n/(n-k)  
sig2 <- c(t(e) %\*% e)/(n-k)  
  
# step 3: coefficients' standard error   
## --- base case: with homoskedastic formula (OLS)  
## --- we have done this before!  
xx <- solve(t(x)%\*%x)  
v0 <- xx\*sig2  
s0 <- sqrt(diag(v0)) # Homoskedastic formula  
  
## --- compare case: with White formula  
u1 <- x\*(e%\*%matrix(1,1,k))  
v1 <- xx %\*% (t(u1)%\*%u1) %\*% xx  
s1 <- sqrt(diag(v1))

* 代表矩阵相乘（也即我们正常理解的矩阵相乘，要求矩阵维度相容）
* 代表矩阵元素相乘（也即两个矩阵的对应元素直接相乘）
* 表示行向量（1行k列）。