

HW2 方法的实证应用——使用 CEPS 数据

吴思思 15320171151911 经济系西方经济学

2019 年 3 月 29 日

本次作业所用数据为“中国教育追踪调查（CEPS）”数据库。中国教育追踪调查（CEPS）是由中国人民大学中国调查与数据中心设计与实施的、具有全国代表性的大型追踪调查项目，旨在揭示家庭、学校、社区以及宏观社会结构对于个人教育产出的影响，并进一步探究教育产出在个人生命历程中发生作用的过程。中国教育追踪调查 (CEPS) 采用 2013-2014 年的基线数据，以初中一年级（7 年级）和初中三年级（9 年级）两个同期群为调查起点，以人口平均受教育水平和流动人口比例为分层变量从全国随机抽取了 28 个县级单位（县、区、市）作为调查点。调查的执行以学校为基础，在入选的县级单位随机抽取了 112 所学校、438 个班级进行调查，被抽中班级的学生全体入样，基线调查共调查了约 2 万名学生。

本次作业采用数据库中对学生问卷调查的数据并使用 R，旨在简要分析认知水平、性别、年级对学生数学成绩的影响。变量说明如下：

变量	变量描述
数学成绩 stamat	学生 2013 年期中考试标准化数学成绩，标准化得分按学校、年级分别计算
认知水平 cog3pl	用三参数的 IRT 模型估计出的学生认知能力测试标准化总分，连续变量
性别 a01	0 女，1 男
年级 grade9	0 七年级，1 九年级

实证结果：

```
> library(haven)
> student <- read_dta("D:/CEPS/student.dta")
> View(student)
> reg1 = lm(stdmat ~ cog3pl, data=student)
> summary(reg1)

Call:
lm(formula = stdmat ~ cog3pl, data = student)

Residuals:
    Min       1Q   Median       3Q      Max
-61.668  -5.528   1.178   6.293  74.044

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  69.96374    0.06691 1045.61  <2e-16 ***
cog3pl        4.32367    0.07780   55.57  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.22 on 18988 degrees of freedom
(497 observations deleted due to missingness)
Multiple R-squared:  0.1399,    Adjusted R-squared:  0.1399
F-statistic: 3088 on 1 and 18988 DF, p-value: < 2.2e-16
```

回归一根据上图实证结果，我们可以得知认知水平对中学生的数学成绩有着显著的影响：认知水平越高，其数学成绩也会越好。

```
>
> reg2 = lm(stdmat ~ cog3pl*a01, data=student)
> summary(reg2)

Call:
lm(formula = stdmat ~ cog3pl * a01, data = student)

Residuals:
    Min       1Q   Median       3Q      Max
-61.129  -5.472   1.185   6.244  73.466

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  68.2320    0.2105 324.141  < 2e-16 ***
cog3pl        4.9182    0.2419  20.330  < 2e-16 ***
a01          1.1880    0.1341   8.860  < 2e-16 ***
cog3pl:a01   -0.4255    0.1564  -2.721  0.00651 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.171 on 18725 degrees of freedom
(758 observations deleted due to missingness)
Multiple R-squared:  0.1432,    Adjusted R-squared:  0.143
F-statistic: 1043 on 3 and 18725 DF, p-value: < 2.2e-16
```

回归二加入性别及认知水平与性别的交互项这两个解释变量后，结果依然显著。但削弱了认知水平对数学成绩的影响。

```
>
> reg3 = lm(stdmat ~ cog3pl*a01*grade9, data=student)
> summary(reg3)

Call:
lm(formula = stdmat ~ cog3pl * a01 * grade9, data = student)

Residuals:
    Min       1Q   Median       3Q      Max
-61.094  -5.498   1.171   6.226  73.416

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    68.27326    0.28796  237.096 < 2e-16 ***
cog3pl         4.24467    0.32676  12.990 < 2e-16 ***
a01            1.18041    0.18497   6.382 1.79e-10 ***
grade9        -0.10176    0.42209  -0.241  0.80949
cog3pl:a01     -0.04203    0.21430  -0.196  0.84451
cog3pl:grade9  1.51254    0.48656   3.109  0.00188 **
a01:grade9      0.02237    0.26864   0.083  0.93364
cog3pl:a01:grade9 -0.85244    0.31382  -2.716  0.00661 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.169 on 18721 degrees of freedom
(758 observations deleted due to missingness)
Multiple R-squared:  0.1437,    Adjusted R-squared:  0.1433
F-statistic: 448.7 on 7 and 18721 DF,  p-value: < 2.2e-16
```

回归三则在回归二的基础上加入了年级，年级与认知水平及性别的交互项以及三者的交互，从结果可以看出九年级相较于七年级的认知水平对中学生的数学成绩有更大的积极影响。但加入性别后这一影响减弱了。