

# 说明文档

## 数据来源

本研究以河南省某县高级中学 2014-2022 年参加高考的学生为研究对象。研究使用的行政数据包括学生模拟考试成绩、高考成绩、个人特征、录取结果及贫困生身份信息。为保证数据质量，我们剔除了同年级重名样本（占比 2.9%），并将研究对象限定为同时参加模拟考试和高考的学生。最终获得有效样本 8496 人。

学生的模拟考试成绩、高考成绩及录取结果均来自高级中学的行政数据。我们对高考成绩和模拟考试成绩按年份、文理科分别进行了标准化处理。依据国家对高校的分类标准，我们将录取院校分为四类：一是“985 工程”和“211 工程”重点大学，二是普通一本院校，三是二本院校，四是高职高专院校。

贫困生身份信息来源于《贫困人口微观追踪数据库》。该数据库由政府部门于 2014 年建立，记录了国家级贫困县内贫困户的家庭和个人信息。在本研究的样本中，贫困生共计 730 人，占总样本的 9.7%。

## 提交程序文件说明

共提交五份数据文件，将使用的程序整理成了 dofile、dofile\_appendix 两个文档，及对应的日志文件和图表生成截图。

表 1 程序文档

程序文档	具体内容
dofile	正文图表的生成程序
dofile_appendix	附录图表的生成程序
日志文件.smcl	运行 master.dofile 生成的日志文件
日志文件_appendix.smcl	运行 master_appendix.dofile 生成的日志文件
图表截图	生成文章图表的截图

本研究程序中使用的数据和出现的变量说明见下表：

表 2 变量说明

使用数据	具体内容
merge2.dta	高考成绩、模考成绩及个人特征
outcome2.dta	录取结果及个人特征
survey2.dta	调查问卷
变量	具体含义
female	表示性别的虚拟变量，女生取值 1
poor	表示贫困生的虚拟变量
stem	表示理科生的虚拟变量
restudy	表示复读生的虚拟变量
local	表示本地户口的虚拟变量
age_at_gaokao	年龄

total_4	模考总分
total_gaokao	高考总分
chinese_4	模考语文分数
math_4	模考数学分数
comprehensive_4	模考综合分数
english_4	模考英语分数
chinese_gaokao	高考语文分数
math_gaokao	高考数学分数
comprehensive_gaokao	高考综合分数
english_gaokao	高考英语分数
dzttotal_4	标准化高考分数-标准化模考分数

---