

Lecture IV: Paper Replication

Stata Training Course 2024 Fall

George Chou

School of Public Finance and Taxation, ZUEL

December 8th, 2024



Outline

- 1 Introduction
- 2 Stylized Fact & Data
- 3 Baseline Regression
- 4 Mechanism Analysis
 - City Attraction
 - University Attraction
- 5 Further Analysis
 - City Living Cost
 - Relative Development Level
 - Heterogeneity
- 6 Appendix

- 1 Introduction
- 2 Stylized Fact & Data
- 3 Baseline Regression
- 4 Mechanism Analysis
 - City Attraction
 - University Attraction
- 5 Further Analysis
 - City Living Cost
 - Relative Development Level
 - Heterogeneity
- 6 Appendix

- 人才是经济高质量发展的动力源泉。中国政府一直高度重视人才培养，积极发挥人才作用。与此同时，地方政府大力引进人才，不断推出各类人才优待政策，各地区“抢人大战”愈演愈烈。人才政策和“抢人大战”的背后，是各级政府认识到人才在社会发展和经济高质量增长中的重要性。
- 已有部分文献对中国人才流动现象进行研究，但既有文献大多考虑的是成熟人才流动，即高技能劳动力与高受教育水平人员，却忽视了中国一个潜在的大规模人才流动现象——高考。

- 首先, 经验证明城市人均 GDP 显著提高城市内高校招生分数线, 即准大学生更倾向于报考位于经济发达地区的院校。
- 其次, 从就业、教育、基础设施与医疗条件等角度分析了城市吸引力这一机制; 从高校师资力量、经费与课题分析了高校吸引力这一机制。
- 最后, 研究发现城市高房价、物价上升与交通拥堵对准大学报考意愿的抑制效应。

- In this lecture, we will use some external command below for the paper replication.

```
1 center // 去中心化处理
2 winsor2 // 缩尾处理
3 binscatter // 分组散点图
4 reghdfe // 高维固定效应回归
5 ivreghdfe // 高维固定效应工具变量回归
6 outreg2 // 回归结果输出
```

- 1 Introduction
- 2 Stylized Fact & Data
- 3 Baseline Regression
- 4 Mechanism Analysis
 - City Attraction
 - University Attraction
- 5 Further Analysis
 - City Living Cost
 - Relative Development Level
 - Heterogeneity
- 6 Appendix

- 图 1 绘制了滞后一期人均 GDP 对数与标准化平均分数线之间的分仓散点图。具体地，按滞后一期人均 GDP 对数大小分为两组，垂直虚线左侧为低组，右侧为高组。可见，全样本、低组与高组，均呈现出滞后一期人均 GDP 对数与分数线之间存在正相关关系。

```
1 use main.dta, clear
2 xtile x = lnLrgdp, n(2)
3 sum lnLrgdp, d
4 binscatter mean_s1 lnLrgdp, ///
5 by(x) n(100) legend(off) ///
6 xline(10.39069, lp(dash_dot)) ///
7 xtitle(滞后一期人均 GDP 对数) ///
8 ytitle(标准化平均分数线) scheme(s1mono)
```

Stylized Fact

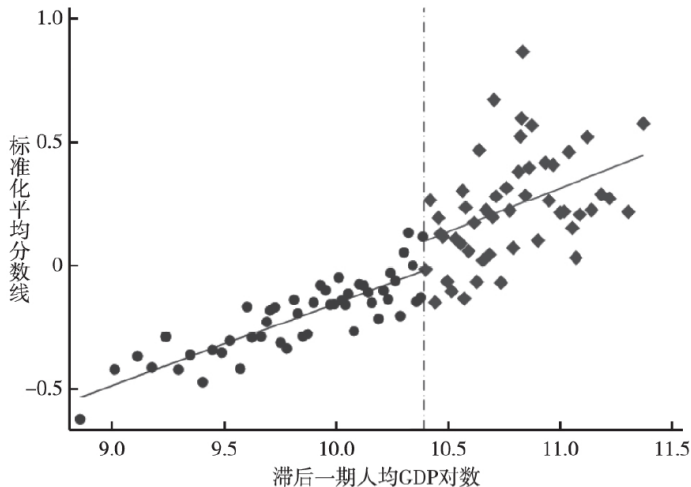


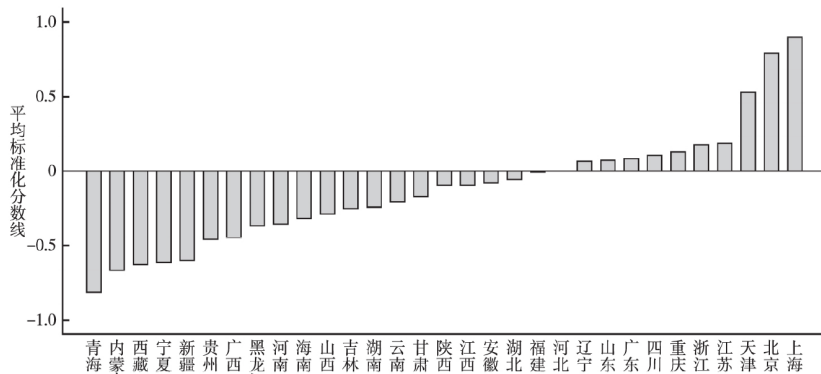
图: 经济发展水平分组的分仓散点图

- 随后，本文将标准化分数线简单平均到省市层面，并绘制柱状图¹，如图 2 所示。可见，分数较高的省市基本位于东部地区，而分数较低的省市则位于中西部地区。分数线的地域分布大致和中国经济的地域分布相合，这初步表明经济发达地区的整体分数线较高，符合“孔雀东南飞”现象。

```
1 bysort province: egen 平均分数线 = mean(mean_s1)
2 duplicates drop province, force
3 graph bar 平均分数线 , over(province, sort(1) ///
4 label(angle(vertical))) ///
5 ytitle(平均分数线) scheme(s1mono)
```

¹原文为直方图，经作者反馈已更正。

Stylized Fact



图：省市平均标准化分数线

- 本文核心数据是高校数据集，包括高校招生分数线数据与高校科技统计数据。本文从新浪高考网搜集了 2006 ~ 2018 年高校招生分数线，包括各年高校在各省各市各科目（文科与理科）、各批次（本科一批、本科二批与本科三批）的平均分与最高分。
- 高校科技统计数据来自教育部发布的 2007 ~ 2017 年《高等学校科技统计资料汇编》（以下简称高校汇编）。高校汇编提供了各高校人员情况、经费情况和科技成果，例如教学与科研人员总数、科技经费总额与专著数量等。
- 由于研究期内存在高校改名等情况，因此本文基于教育部发布的《2020 年全国普通高等学校名单》（以下简称高校名单），辅以网络手工收集等方式，确定高校唯一标识码。此外，高校名单还提供了高校主管部门、所在地与办学性质信息。通过高校唯一标识码与年份，将高校招生分数线数据和高校科技统计数据进行合并，形成一套完整的高校研究数据集。

Descriptive Statistics

变量	含义	样本量	均值	方差	最小值	最大值
<i>mean_s</i>	标准化平均分数线	424,552	0.014	0.997	-11.491	12.182
<i>lnLrgdp</i>	滞后一期人均 GDP 对数	424,552	10.316	0.569	8.772	11.374
<i>lnLthird</i>	滞后一期第三产业增加值份额对数	424,552	3.818	0.254	2.303	4.390
<i>lnLtea</i>	滞后一期高校老师数量对数	424,552	9.521	1.290	0	11.152
<i>Lhighrail</i>	滞后一期高铁开通虚拟变量	424,552	0.518	0.500	0	1
<i>Ltemp</i>	滞后一期城市年平均气温	424,552	14.466	4.502	-1.085	26.376
<i>Lrain</i>	滞后一期城市年降水量	424,552	984.439	440.813	209.523	2603.182
<i>Lflood</i>	滞后一期城市非重复洪水次数	424,552	0.178	0.768	0	9

图：主要变量的描述性统计

- 1 Introduction
- 2 Stylized Fact & Data
- 3 Baseline Regression**
- 4 Mechanism Analysis
 - City Attraction
 - University Attraction
- 5 Further Analysis
 - City Living Cost
 - Relative Development Level
 - Heterogeneity
- 6 Appendix

- 本文旨在探究准大学生在报考时是否倾向于选择经济发达地区的院校，即地区经济发展水平是否显著提升了区域内高校招生分数线。
- 故本文选取高校平均分数线作为被解释变量，并作标准化处理：（各下标中， i 代表高校， p 代表招生省市， s 代表科目， b 代表批次， t 代表年份。）

$$mean_s_{ipsbt} = \frac{[mean_{ipsbt} - average(mean_{ipsbt})]}{std(mean_{ipsbt})}$$

- 由于各省高考在每年年中开始，最迟到 9 月初完成所有流程，因此考生在报考时考虑的是往期各院校所在城市的经济发展水平。对此，本文核心解释变量选取城市人均 GDP 作为地区经济发展水平的替代，并取对数以及滞后一期 ($\ln Lrgdp$)。

- 本文因果识别面临的内生性问题中的互为因果较弱，主要是遗漏变量与测量误差。尤其是难以控制同时影响考生选择和城市经济发展的不可观测因素，例如城市文化氛围、环境气候与对待人才态度等。
- 对此，本文选取一个地理环境变量——城市洪水，作为城市经济发展水平的工具变量。
 - 《水情年报》定义的洪水对准大学生影响较小，这保证了工具变量的外生性。
 - 洪水通过抑制地方经济活动（例如物流运输）从而对经济发展水平产生负面影响，进而影响准大学生选择，从而满足工具变量要求。

- 滞后一期城市产业结构，用第三产业增加值份额衡量，取对数 ($\ln Lthird$)。
- 滞后一期城市高等教育资源丰度，用城市高校老师数代替，取对数 ($\ln Ltea$)。
- 滞后一期城市是否开通高铁的虚拟变量 ($Lhighrail$)。
- 城市气候特征，包括城市年平均气温 ($Ltemp$) 与年降水量 ($Lrain$)。
- 城市统计数据来自 2005 ~ 2018 年《中国城市统计年鉴》。城市房价与 CPI 数据来自 CEIC 数据库。城市气候数据来自中国气象科学数据中心。水情数据来自水利部发布的历年《水情年报》。对于部分缺失数据，本文使用前后期平均值进行了补缺。相关变量以 2000 年为基期进行了平减处理。

- OLS Model:

$$mean_s_{cipsbt} = \alpha + \beta \ln Lrgdp_{ct} + \gamma X + F + \varepsilon_{cit}$$

- X : Controls.
- $F = \delta_i + \mu_t + \tau_{cp}$: Fixed effects.²
- ε_{cit} : Random errors.
- Generally, regressions are clustered at the *city level* where the university is located.

²原文 δ_i 的下标为 t ，经作者反馈已更正。

- 1st-stage Model:

$$\ln Lrgdp_{ct} = \alpha_1 + \beta_1 \ln Lflood_{ct} + \gamma_1 X + F + \varepsilon_{cit}^I$$

- 2nd-stage Model:

$$mean_s_{cipsbt} = \alpha_2 + \beta_2 \ln \widehat{Lrgdp}_{ct} + \gamma_2 X + F + \varepsilon_{cit}^{II}$$

- Weak instrument test:
 - Stock Yogo test: All Kleibergen-Paap rk Wald F statistics > 16.38 .
 - All 1st-stage F values > 10 .

- Replacing core variables.
 - 被解释变量替换为城市内高校最高分数线。
 - 核心解释变量替换为市辖区滞后一期人均 GDP 对数。
 - 剔除北京、上海两个头部城市的样本。
- Adding more controls.
 - 加入地方政府对人才重视程度³⁴作为控制变量。
 - 加入额外的城市层面控制变量，包括城市经济规模（滞后一期 GDP 对数）与滞后一期是否开通机场虚拟变量。
 - 加入高校层面滞后一期控制变量，包括高校专著数对数、论文数对数与奖项数对数。
- Adjusting data cleaning steps.

³原文：采用基于城市政府工作报告的词频分析方法，计算人才与创新关键词数量占工作报告总词数的比例。

⁴经作者反馈更正：经检查，当时应该没有纳入“创新”关键词。为了防止人为选择的干扰，我重新构建了两个词频变量：第一个不加任何选择，直接计算词频；第二个按“人才”、“创”（以纳入“原创”、“双创”）和“知识产权”刷选关键词。最终结果稳健。

- 1 Introduction
- 2 Stylized Fact & Data
- 3 Baseline Regression
- 4 Mechanism Analysis**
 - City Attraction
 - University Attraction
- 5 Further Analysis
 - City Living Cost
 - Relative Development Level
 - Heterogeneity
- 6 Appendix

- 1 Introduction
- 2 Stylized Fact & Data
- 3 Baseline Regression
- 4 Mechanism Analysis
 - City Attraction
 - University Attraction
- 5 Further Analysis
 - City Living Cost
 - Relative Development Level
 - Heterogeneity
- 6 Appendix

- 从考生角度来看，高经济发展水平可能意味着城市在工作、教育、福利与设施、医疗，以及生活消费水平方面更具优势。尤其对于毕业后打算在目标城市定居的考生，更倾向于选择有着更多优质工作岗位、良好教育水平、完善社会福利与设施建设、充分医疗保障，以及较高生活消费水平的城市。同时除以上因素之外，高经济发展水平也可能意味着较高的城市服务水平。

- 本文选取相关代理变量，用以衡量城市工作、教育、福利设施、医疗与生活消费水平，并替换为基准回归模型中的被解释变量。（均作对数和滞后一期处理，某些变量还做了平减化处理。）
 - 职工平均工资（衡量就业薪资水平）；
 - 人均政府财政教育支出（衡量政府教育投入）；
 - 人均政府财政支出（衡量社会福利水平）；
 - 人均医生数（衡量医疗服务水平）；
 - 人均社会消费品零售额（衡量生活消费水平）；
 - 城市服务业从业人数占城市总人口比例（衡量服务水平与产业结构）。

- 1 Introduction
- 2 Stylized Fact & Data
- 3 Baseline Regression
- 4 Mechanism Analysis
 - City Attraction
 - University Attraction
- 5 Further Analysis
 - City Living Cost
 - Relative Development Level
 - Heterogeneity
- 6 Appendix

- 城市经济发展水平在提高城市吸引力的同时，也可能提高城市内高校的吸引力。例如城市经济越发达，高校经费越多，师资规模越大。高经济发展水平带来的高师资规模，一方面可能由于高校有能力提供更多的教学与科研岗位；另一方面，也可能由于高校老师倾向于选择经济发达城市的高校就职。师资规模是高校实力与吸引力的重要体现，准大学生可能倾向于选择师资雄厚，尤其是有着更多高级职称人员的高校。
- 除了师资规模以外，高校经费与课题也是高校实力的重要一环。一方面，城市经济越发达，政府为高校拨款的财政基础越雄厚，相关政府部门与企业的课题需求越旺盛。另一方面，高校也可以通过与地方企业合作等方式，筹措更多的经费，获得更多的课题。

- 本文从 2007 ~ 2017 年《高等学校科技统计资料汇编》中整理出高校科技人员信息，并替换为基准回归模型中的被解释变量。（均作对数和滞后一期处理。）
 - 教学与科研人员总数（衡量师资规模）；
 - 其中：具有高级职称的科学家与工程师人数 a （衡量高级职称）；
 - 科技经费总额（衡量高校经费）；
 - 科技课题总数（衡量课题数量）；
 - 科技课题当年拨入经费（衡量课题经费）。

- 1 Introduction
- 2 Stylized Fact & Data
- 3 Baseline Regression
- 4 Mechanism Analysis
 - City Attraction
 - University Attraction
- 5 Further Analysis
 - City Living Cost
 - Relative Development Level
 - Heterogeneity
- 6 Appendix

- 1 Introduction
- 2 Stylized Fact & Data
- 3 Baseline Regression
- 4 Mechanism Analysis
 - City Attraction
 - University Attraction
- 5 Further Analysis
 - City Living Cost
 - Relative Development Level
 - Heterogeneity
- 6 Appendix

- 前文研究表明，高经济发展水平能吸引准大学生报考城市内高校。但城市经济越发达，居住成本可能越高昂，代表是大城市的房价、物价与交通拥堵。
- 已有研究发现房价会挤出劳动力，尤其是高技能水平劳动力。本文将 CEIC 数据库提供的城市房价对数与 $\ln Lrgdp$ 交乘，加入回归。
 - 房屋销售价格 $Lprice1$;
 - 房屋销售价格（住宅） $Lprice2$;
 - 商品房销售额与商品房销售面积之比 $Lprice3$;
 - 以及商品房销售额与商品房销售面积之比（住宅） $Lprice4$ 。

- 物价也可能会产生挤出效应。对此，本文将 CEIC 数据库提供的城市 CPI（上一年 = 100）数据转化为 2009 年为基期的定基指数 $Lcpi_{city}$ ，与 $\ln Lrgdp$ 交乘，加入回归。⁵
- 对于交通拥堵，本文从 2015 ~ 2017 年高德地图发布的《中国主要城市交通分析报告》中提取城市年度高峰拥堵平均速度数据，取对数后与 $\ln Lrgdp$ 交乘，加入回归。⁶（高峰拥堵平均速度越大，意味着城市在交通高峰情况下的拥堵成本越低。）

⁵由于城市 CPI 数据缺失严重，大部分地级市从 2010 年才开始有 CPI（上一年 = 100）数据，综合考虑下，本文计算了以 2009 年为基期的 CPI 定基指数。

⁶2015 年是本文所能追寻到的最早数据年份。该报告 2015 年共 45 个城市，2016 年共 60 个城市，2017 年共 100 个城市。

- 1 Introduction
- 2 Stylized Fact & Data
- 3 Baseline Regression
- 4 Mechanism Analysis
 - City Attraction
 - University Attraction
- 5 Further Analysis
 - City Living Cost
 - Relative Development Level
 - Heterogeneity
- 6 Appendix

- 由前文分析可知，城市经济发展水平显著提高了城市内高校的招生分数线，其机制在于城市经济发展水平越高，城市吸引力与高校吸引力越强，但以上都基于城市绝对经济水平。那么，地区间的相对经济发展水平是否也会产生类似现象？例如，高校所在城市相对于招生地的经济越发达，其招生分数线是否也越高？
- 本文定义高校所在地与招生地的相对经济发展水平为前者人均 GDP 与后者人均 GDP 的比值，取对数。
- 从地区相对经济发展水平这一角度进行分析，其优点在于可以解剖考生在报考时的相对心理价值比较。考生可能会对比高校所在城市与自己家乡的相对经济发展水平，也有可能对比两个省市之间的相对经济发展水平，甚至对比两个省会。

- 考生可能对比的高校地：
 - 高校所在城市；
 - 高校所在省省会（直辖市即自身）；
 - 高校所在省内人均 GDP 最高的城市；
 - 高校所在省。
- 考生可能对比的家乡地：
 - 招生省省会（直辖市即自身）；
 - 招生省省内人均 GDP 最高的城市；
 - 招生省市。

- 1 Introduction
- 2 Stylized Fact & Data
- 3 Baseline Regression
- 4 Mechanism Analysis
 - City Attraction
 - University Attraction
- 5 Further Analysis
 - City Living Cost
 - Relative Development Level
 - Heterogeneity
- 6 Appendix

- 招生分组：
 - 招生科目（文科 & 理科）；
 - 招生批次。
- 高校分组：
 - 高校办学类型（公办 & 民办）；
 - 院校归属类型（主体高校 & 独立学院）；
 - 高校主管部门。
- 地理位置分组：（高校所在省市与招生省市）
 - 同一省份；
 - 相邻省份；
 - 距离 1000 公里。⁷

⁷高校所在城市与招生省市间距离为城市与招生省会（直辖市）距离。

- 城市经济发展水平，即人均 GDP 显著提高了城市内高校招生分数线；这意味着准大学生更愿意报考经济发达地区的院校，该结论在内生性处理与多种稳健性检验下均成立。
- 城市经济发展水平通过提高其在工作、教育、医疗、社会福利与设施，以及服务业发展水平等方面的吸引力，进而提升高校招生分数线。经济发达地区的院校在师资、经费与课题方面更具优势，进而提高了高校吸引力。
- 城市房价、物价与交通拥堵会抑制经济发展水平对考生的吸引力。目标城市相较于考生所在地的相对经济发展水平，也会提升高校招生分数线。

- 1 Introduction
- 2 Stylized Fact & Data
- 3 Baseline Regression
- 4 Mechanism Analysis
 - City Attraction
 - University Attraction
- 5 Further Analysis
 - City Living Cost
 - Relative Development Level
 - Heterogeneity
- 6 Appendix

Data Cleansing of “城市人才词频”

```
1 use 地级市基础研究&人才.dta, clear
2 tab keyword
3 gen 人才 = strmatch (keyword, "*人才*")
4 gen 创新 = strmatch (keyword, "*创*") // 原创、双创等
5 gen 知识产权 = strmatch (keyword, "*知识产权*")
```

```
1 bysort city year: egen x = sum(count)
2 gen cipin1 = x/word_sum // 不加筛选
3 gen y = (人才 == 1 | 创新 == 1 | 知识产权 == 1)
4 bysort city year: egen z = sum(count*y)
5 gen cipin2 = z/word_sum // 筛选人才、创新、知识产权
```

Data Cleansing of “城市人才词频”

```
1 keep city year cipun1 cipun2
2 duplicates drop city year, force
3 egen cityid = group(city)
4 xtset cityid year
5 gen Lcipin1 = L.cipun1
6 gen Lcipin2 = L.cipun2
7 drop cityid
8 save 城市人才词频 - 更新.dta, replace
```

Recommendation

- Angrist, Joshua D., and Jörn-Steffen Pischke. “Mostly Harmless Econometrics: An Empiricist’s Companion.” *Princeton University Press*, 2009.
- Cunningham, Scott. “Causal Inference: The Mixtape.” *Yale University Press*, 2021.
- Acock, Alan C. “A Gentle Introduction to Stata.” *Stata Press*, 2023.
- 陈强. “计量经济学及 Stata 应用.” 高等教育出版社, 2023.
- 黄炜, 张子尧, 刘安然. “从双重差分法到事件研究法.” 产业经济评论, 2022.
- 张子尧, 黄炜. “事件研究法的实现、问题和拓展.” 数量经济技术经济研究, 2023.

Questions are welcome.