# Stata Exercises

*Due Nov. 18, 2026*

## 第一题：工具变量

2024 年诺贝尔经济学奖授予达龙 · 阿西莫格鲁（Daron Acemoglu）、西蒙 · 约翰逊（Simon Johnson）和詹姆斯 · A · 罗宾逊（James A. Robinson），以表彰他们在“制度如何形成及其对经济繁荣影响”方面的研究。他们在 2001 年发表在经济学顶刊 *AER* 上的文章探讨了**制度是否是人均收入跨国差异的根本原因？**直接回归制度和收入面临严重的内生性：

- 反向因果：富国可能选择更好的制度；
- 遗漏变量：地理、文化、疾病等可能同时影响制度和收入；
- 测量误差：制度变量存在测量误差。

AJR (2001) 的核心识别策略是：欧洲殖民者在不同殖民地面临的**潜在定居者死亡率**不同。在死亡率高的地方，欧洲人难以大规模定居，更有可能建立“汲取型/攫取型/掠夺型”（extractive）制度，只保护殖民者掠夺资源，不保护私有产权、不约束政府；在死亡率低、适合定居的地方，欧洲移民更可能复制母国制度，建立“包容型”（inclusive）制度，强调产权和法治；而殖民制度在独立后具有持续性。因此，殖民者死亡率通过“定居可能性 → 早期制度 → 当代制度”影响当代人均收入，而它本身不直接进入当代收入方程。AJR 用 17–19 世纪士兵、主教、水手的死亡率作为当代制度质量（Political Risk Services：征用风险保护指数，0–10，越高越好）的**工具变量**，用 2SLS 估计制度对人均收入的因果效应。

**理论链条：**

$$\text{潜在定居者死亡率} \Rightarrow \text{殖民活动} \Rightarrow \text{早期制度} \Rightarrow \text{当代制度} \Rightarrow \text{当代经济表现}$$

本练习要求你用 Stata 复现 AJR (2001) 的部分关键内容，理解工具变量的相关条件、外生性/排除约束，以及 2SLS、弱 IV、过度识别约束和内生性检验。

# 基准回归方程

$$\log y_i = \mu + \alpha R_i + \mathbf{X}_i'\gamma + \varepsilon_i \tag{1}$$

- 因变量是人均收入对数。
- $R_i$ 是当前制度质量。
- $\mathbf{X}_i$ 是控制变量项，例如纬度、大陆虚拟变量等。

表 2 用 OLS 估计式 (1)，表 4 用 2SLS 估计式 (1)，其中 $\log M_i$ 是 $R_i$ 的工具变量。

# 第一阶段方程

$$R_i = \zeta + \beta \log M_i + \mathbf{X}_i'\delta + v_i \tag{5}$$

TABLE 2—OLS REGRESSIONS

| | Whole world (1) | Base sample (2) | Whole world (3) | Whole world (4) | Base sample (5) | Base sample (6) | Whole world (7) | Base sample (8) |
|---|---|---|---|---|---|---|---|---|
| | Dependent variable is log GDP per capita in 1995 | | | | | | Dependent variable is log output per worker in 1988 | |
| Average protection against expropriation risk, 1985–1995 | 0.54 (0.04) | 0.52 (0.06) | 0.47 (0.06) | 0.43 (0.05) | 0.47 (0.06) | 0.41 (0.06) | 0.45 (0.04) | 0.46 (0.06) |
| Latitude | | | 0.89 (0.49) | 0.37 (0.51) | 1.60 (0.70) | 0.92 (0.63) | | |
| Asia dummy | | | | −0.62 (0.19) | | −0.60 (0.23) | | |
| Africa dummy | | | | −1.00 (0.15) | | −0.90 (0.17) | | |
| "Other" continent dummy | | | | −0.25 (0.20) | | −0.04 (0.32) | | |
| $R^2$ | 0.62 | 0.54 | 0.63 | 0.73 | 0.56 | 0.69 | 0.55 | 0.49 |
| Number of observations | 110 | 64 | 110 | 110 | 64 | 64 | 108 | 61 |

*Notes:* Dependent variable: columns (1)–(6), log GDP per capita (PPP basis) in 1995, current prices (from the World Bank's World Development Indicators 1999); columns (7)–(8), log output per worker in 1988 from Hall and Jones (1999). Average protection against expropriation risk is measured on a scale from 0 to 10, where a higher score means more protection against expropriation, averaged over 1985 to 1995, from Political Risk Services. Standard errors are in parentheses. In regressions with continent dummies, the dummy for America is omitted. See Appendix Table A1 for more detailed variable definitions and sources. Of the countries in our base sample, Hall and Jones do not report output per worker in the Bahamas, Ethiopia, and Vietnam.

# 制度决定链方程

$$R_i = \lambda_R + \beta_R C_i + \mathbf{X}_i'\gamma_R + \nu_{Ri} \tag{2}$$

$$C_i = \lambda_C + \beta_C S_i + \mathbf{X}_i'\gamma_C + \nu_{Ci} \tag{3}$$

$$S_i = \lambda_S + \beta_S \log M_i + \mathbf{X}_i'\gamma_S + \nu_{Si} \tag{4}$$

- 式 (2)：当前制度 $R$ 取决于早期制度 $C$。$\beta_R$ 度量早期制度对当前制度的持续性。
- 式 (3)：早期制度 $C$ 取决于欧洲人定居程度 $S$。定居者越多，越可能建立产权保护和制衡制度，故预期 $\beta_C > 0$。
- 式 (4)：定居程度 $S$ 取决于殖民者死亡率对数 $\log M$。死亡率越高，定居越少，故预期 $\beta_S < 0$。

核心识别公式可概括为：

$$\underbrace{\log M_i}_{\text{工具变量}} \Rightarrow \underbrace{S_i}_{\text{定居}} \Rightarrow \underbrace{C_i}_{\text{早期制度}} \Rightarrow \underbrace{R_i}_{\text{当前制度}} \Rightarrow \underbrace{\log y_i}_{\text{当前收入}}$$

## Questions:

1. 表 3 的结果如何支持上述核心渠道，每个环节对应什么系数?
2. 复现表 4 Panel A 的 (1)–(9) 列，将所有结果合成一张表。
3. 表 4 第 (1) 列与 OLS 系数相比如何变化? 解释第 (1) 列的 2SLS 系数含义。
4. 工具变量需要满足哪两个核心条件? 本文用什么经济逻辑支持这两个条件?
5. 弱工具变量会导致什么后果? 结合表 4 判断是否存在弱工具变量问题。
6. 表 4 列 (7)–(8) 加入大陆虚拟变量（以美洲为基准）。非洲虚拟变量约-0.46，不显著；而表 2 中非洲虚拟变量显著为负。这是否足以证明“非洲更穷是因为制度更差”?
7. 为什么作者不直接把 $S_i$ 或 $C_i$ 作为主工具变量? 表 8 是如何检验“殖民者死亡率只通过制度影响收入”的（排除性约束）?

TABLE 3—DETERMINANTS OF INSTITUTIONS

| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) |
|---|---|---|---|---|---|---|---|---|---|---|
| Panel A | Dependent Variable Is Average Protection Against Expropriation Risk in 1985–1995 | | | | | | | | | |
| Constraint on executive in 1900 | 0.32 | 0.26 | | | | | | | | |
| | (0.08) | (0.09) | | | | | | | | |
| Democracy in 1900 | | | 0.24 | 0.21 | | | | | | |
| | | | (0.06) | (0.07) | | | | | | |
| Constraint on executive in first year of independence | | | | | 0.25 | 0.22 | | | | |
| | | | | | (0.08) | (0.08) | | | | |
| European settlements in 1900 | | | | | | | 3.20 | 3.00 | | |
| | | | | | | | (0.61) | (0.78) | | |
| Log European settler mortality | | | | | | | | | −0.61 | −0.51 |
| | | | | | | | | | (0.13) | (0.14) |
| Latitude | | 2.20 | | 1.60 | | 2.70 | | 0.58 | | 2.00 |
| | | (1.40) | | (1.50) | | (1.40) | | (1.51) | | (1.34) |
| $R^2$ | 0.2 | 0.23 | 0.24 | 0.25 | 0.19 | 0.24 | 0.3 | 0.3 | 0.27 | 0.3 |
| Number of observations | 63 | 63 | 62 | 62 | 63 | 63 | 66 | 66 | 64 | 64 |

| Panel B | Dependent Variable Is Constraint on Executive in 1900 | | | | Dependent Variable Is Democracy in 1900 | | | | Dependent Variable Is European Settlements in 1900 | |
|---|---|---|---|---|---|---|---|---|---|---|
| European settlements in 1900 | 5.50 | 5.40 | | | 8.60 | 8.10 | | | | |
| | (0.73) | (0.93) | | | (0.90) | (1.20) | | | | |
| Log European settler mortality | | | −0.82 | −0.65 | | | −1.22 | −0.88 | −0.11 | −0.07 |
| | | | (0.17) | (0.18) | | | (0.24) | (0.25) | (0.02) | (0.02) |
| Latitude | | 0.33 | | 3.60 | | 1.60 | | 7.60 | | 0.87 |
| | | (1.80) | | (1.70) | | (2.30) | | (2.40) | | (0.19) |
| $R^2$ | 0.46 | 0.46 | 0.25 | 0.29 | 0.57 | 0.57 | 0.28 | 0.37 | 0.31 | 0.47 |
| Number of observations | 70 | 70 | 75 | 75 | 67 | 67 | 68 | 68 | 73 | 73 |

*Notes:* All regressions are OLS. Standard errors are in parentheses. Regressions with constraint on executive in first year of independence also include years since independence as a regressor. Average protection against expropriation risk is on a scale from 0 to 10, where a higher score means more protection against expropriation of private investment by government, averaged over 1985 to 1995. Constraint on executive in 1900 is on a scale from 1 to 7, with a higher score indicating more constraints. Democracy in 1900 is on a scale from 0 to 10, with a higher score indicating more democracy. European settlements is percent of population that was European or of European descent in 1900. See Appendix Table A1 for more detailed variable definitions and sources.

TABLE 4—IV REGRESSIONS OF LOG GDP PER CAPITA

| | Base sample (1) | Base sample (2) | Base sample without Neo-Europes (3) | Base sample without Neo-Europes (4) | Base sample without Africa (5) | Base sample without Africa (6) | Base sample with continent dummies (7) | Base sample with continent dummies (8) | Base sample, dependent variable is log output per worker (9) |
|---|---|---|---|---|---|---|---|---|---|
| **Panel A: Two-Stage Least Squares** | | | | | | | | | |
| Average protection against expropriation risk 1985–1995 | 0.94 | 1.00 | 1.28 | 1.21 | 0.58 | 0.58 | 0.98 | 1.10 | 0.98 |
| | (0.16) | (0.22) | (0.36) | (0.35) | (0.10) | (0.12) | (0.30) | (0.46) | (0.17) |
| Latitude | | −0.65 | | 0.94 | | 0.04 | | −1.20 | |
| | | (1.34) | | (1.46) | | (0.84) | | (1.8) | |
| Asia dummy | | | | | | | −0.92 | −1.10 | |
| | | | | | | | (0.40) | (0.52) | |
| Africa dummy | | | | | | | −0.46 | −0.44 | |
| | | | | | | | (0.36) | (0.42) | |
| "Other" continent dummy | | | | | | | −0.94 | −0.99 | |
| | | | | | | | (0.85) | (1.0) | |
| **Panel B: First Stage for Average Protection Against Expropriation Risk in 1985–1995** | | | | | | | | | |
| Log European settler mortality | −0.61 | −0.51 | −0.39 | −0.39 | −1.20 | −1.10 | −0.43 | −0.34 | −0.63 |
| | (0.13) | (0.14) | (0.13) | (0.14) | (0.22) | (0.24) | (0.17) | (0.18) | (0.13) |
| Latitude | | 2.00 | | −0.11 | | 0.99 | | 2.00 | |
| | | (1.34) | | (1.50) | | (1.43) | | (1.40) | |
| Asia dummy | | | | | | | 0.33 | 0.47 | |
| | | | | | | | (0.49) | (0.50) | |
| Africa dummy | | | | | | | −0.27 | −0.26 | |
| | | | | | | | (0.41) | (0.41) | |
| "Other" continent dummy | | | | | | | 1.24 | 1.1 | |
| | | | | | | | (0.84) | (0.84) | |
| $R^2$ | 0.27 | 0.30 | 0.13 | 0.13 | 0.47 | 0.47 | 0.30 | 0.33 | 0.28 |
| **Panel C: Ordinary Least Squares** | | | | | | | | | |
| Average protection against expropriation risk 1985–1995 | 0.52 | 0.47 | 0.49 | 0.47 | 0.48 | 0.47 | 0.42 | 0.40 | 0.46 |
| | (0.06) | (0.06) | (0.08) | (0.07) | (0.07) | (0.07) | (0.06) | (0.06) | (0.06) |
| Number of observations | 64 | 64 | 60 | 60 | 37 | 37 | 64 | 64 | 61 |

*Notes:* The dependent variable in columns (1)–(8) is log GDP per capita in 1995, PPP basis. The dependent variable in column (9) is log output per worker, from Hall and Jones (1999). "Average protection against expropriation risk 1985–1995" is measured on a scale from 0 to 10, where a higher score means more protection against risk of expropriation of investment by the government, from Political Risk Services. Panel A reports the two-stage least-squares estimates, instrumenting for protection against expropriation risk using log settler mortality; Panel B reports the corresponding first stage. Panel C reports the coefficient from an OLS regression of the dependent variable against average protection against expropriation risk. Standard errors are in parentheses. In regressions with continent dummies, the dummy for America is omitted. See Appendix Table A1 for more detailed variable descriptions and sources.

TABLE 8—OVERIDENTIFICATION TESTS

| | Base sample (1) | Base sample (2) | Base sample (3) | Base sample (4) | Base sample (5) | Base sample (6) | Base sample (7) | Base sample (8) | Base sample (9) | Base sample (10) |
|---|---|---|---|---|---|---|---|---|---|---|
| **Panel A: Two-Stage Least Squares** | | | | | | | | | | |
| Average protection against expropriation risk, 1985–1995 | 0.87 | 0.92 | 0.71 | 0.68 | 0.72 | 0.69 | 0.60 | 0.61 | 0.55 | 0.56 |
| | (0.14) | (0.20) | (0.15) | (0.20) | (0.14) | (0.19) | (0.14) | (0.17) | (0.12) | (0.14) |
| Latitude | | −0.47 | | −0.34 | | 0.31 | | −0.41 | | −0.16 |
| | | (1.20) | | (1.10) | | (1.05) | | (0.92) | | (0.81) |
| **Panel B: First Stage for Average Protection Against Expropriation Risk** | | | | | | | | | | |
| European settlements in 1900 | 3.20 | 2.90 | | | | | | | | |
| | (0.62) | (0.83) | | | | | | | | |
| Constraint on executive in 1900 | | | 0.32 | 0.26 | | | | | | |
| | | | (0.08) | (0.09) | | | | | | |
| Democracy in 1900 | | | | | 0.24 | 0.20 | | | | |
| | | | | | (0.06) | (0.07) | | | | |
| Constraint on executive in first year of independence | | | | | | | 0.25 | 0.22 | | |
| | | | | | | | (0.08) | (0.08) | | |
| Democracy in first year of independence | | | | | | | | | 0.19 | 0.17 |
| | | | | | | | | | (0.05) | (0.05) |
| $R^2$ | 0.30 | 0.30 | 0.20 | 0.24 | 0.24 | 0.26 | 0.19 | 0.25 | 0.26 | 0.30 |
| **Panel C: Results from Overidentification Test** | | | | | | | | | | |
| $p$-value (from chi-squared test) | [0.67] | [0.96] | [0.09] | [0.20] | [0.11] | [0.28] | [0.67] | [0.79] | [0.22] | [0.26] |
| **Panel D: Second Stage with Log Mortality as Exogenous Variable** | | | | | | | | | | |
| Average protection against expropriation risk, 1985–1995 | 0.81 | 0.88 | 0.45 | 0.42 | 0.52 | 0.48 | 0.49 | 0.49 | 0.4 | 0.41 |
| | (0.23) | (0.30) | (0.25) | (0.30) | (0.23) | (0.28) | (0.23) | (0.25) | (0.18) | (0.19) |
| Log European settler mortality | −0.07 | −0.05 | −0.25 | −0.26 | −0.21 | −0.22 | −0.14 | −0.14 | −0.19 | −0.19 |
| | (0.17) | (0.18) | (0.16) | (0.17) | (0.15) | (0.16) | (0.16) | (0.15) | (0.13) | (0.12) |
| Latitude | | −0.52 | | 0.38 | | 0.28 | | −0.38 | | −0.17 |
| | | (1.15) | | (0.89) | | (0.86) | | (0.84) | | (0.73) |

*Notes:* Panel A reports the two-stage least-squares estimates with log GDP per capita (PPP basis) in 1995 as the dependent variable, and Panel B reports the corresponding first stage (latitude is included in even-numbered columns but is never significant and not reported here to save space). Panel C reports the $p$-value for the null hypothesis that the coefficient on average protection against expropriation risk in the second-stage regression (i.e., Panel A) is the same as when instrumented using log mortality of settlers in addition to the indicated instruments. Panel D reports results from the regression in which log mortality is included as an exogenous variable and current institutions are instrumented using the alternative instrument indicated. Standard errors are in parentheses. All regressions with constraint on executive and democracy in first year of independence also include years since independence as a regressor. All regressions have 60 observations, except those with democracy in 1900 which have 59 observations and those with European settlements in 1900 which have 63 observations.