Introductory Econometrics

Endogeneity and Instrument Variable Estimation

Yaohan Chen

School of Big Data and Statistics, Anhui University

Spring, 2025

Endogeneity

- We say that there is *endogeneity* in the linear regression model

$$Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + u_i$$

  if $\boldsymbol{\beta}$ is the parameter of interest and

$$\mathrm{E}\left(\boldsymbol{X}_i u_i\right) \neq 0.$$

- Sources of endogeneity.
  - Omitted Variable Bias
  - Simultaneous Equations Bias
  - Measurement error in regressors

Endogeneity Issue: Omitted Variable Bias

- Suppose the true regression model is

$$Y_i = \boldsymbol{X}_{i1}'\boldsymbol{\beta}_1 + \boldsymbol{X}_{i2}'\boldsymbol{\beta}_2 + u_i$$

  where $\boldsymbol{X}_{i1}$ and $\boldsymbol{X}_{i2}$ are $k_1 \times 1$ and $k_2 \times 1$ vectors respectively.
- If we omit $\boldsymbol{X}_{i2}$ then OLS estimator for $\boldsymbol{\beta}_1$ is biased and inconsistent.

Endogeneity Issue: Omitted Variable Bias

- Suppose the true regression model is

$$Y_i = \boldsymbol{X}'_{i1}\boldsymbol{\beta}_1 + \boldsymbol{X}'_{i2}\boldsymbol{\beta}_2 + u_i$$

  where $\boldsymbol{X}_{i1}$ and $\boldsymbol{X}_{i2}$ are $k_1 \times 1$ and $k_2 \times 1$ vectors respectively.
- If we omit $\boldsymbol{X}_{i2}$ then OLS estimator for $\boldsymbol{\beta}_1$ is biased and inconsistent.
- This because $\boldsymbol{X} = \left[\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}\right]$

$$\begin{aligned}
\tilde{\boldsymbol{\beta}}_1 &= \left(\boldsymbol{X}^{(1)\prime}\boldsymbol{X}^{(1)}\right)^{-1}\boldsymbol{X}^{(1)\prime}\boldsymbol{Y} \\
&= \boldsymbol{\beta}_1 + \left(\boldsymbol{X}^{(1)\prime}\boldsymbol{X}^{(1)}\right)^{-1}\boldsymbol{X}^{(1)\prime}\boldsymbol{X}^{(2)}\boldsymbol{\beta}_2 + \left(\boldsymbol{X}^{(1)\prime}\boldsymbol{X}^{(1)}\right)^{-1}\boldsymbol{X}^{(1)\prime}\boldsymbol{u}.
\end{aligned}$$

Endogeneity Issue: Measurement Error in Regressors

- Consider the following "true" regression model

$$Y_i = \widetilde{\boldsymbol{X}}_i'\boldsymbol{\beta} + u_i$$

  where $\mathrm{E}\left(\widetilde{\boldsymbol{X}}_i u_i\right) = 0$. But $\widetilde{\boldsymbol{X}}_i$ is not measured accurately or cannot be directly measured.
- Instead, we observe $\boldsymbol{X}_i = \widetilde{\boldsymbol{X}}_i + \boldsymbol{V}_i$, where $\boldsymbol{V}_i$ denotes a $k \times 1$ vector collecting measurement errors. Therefore,

$$Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \underbrace{(u_i - \boldsymbol{V}_i'\boldsymbol{\beta})}_{\varepsilon_i}$$

Endogeneity Issue: Measurement Error in Regressors

- Let $\boldsymbol{X} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n)'$, $\boldsymbol{U} = (u_1, \cdots, u_n)'$, $\boldsymbol{V} = (\boldsymbol{V}_1, \cdots, \boldsymbol{V}_n)'$, $\boldsymbol{Q} = \mathrm{E}\left(\widetilde{\boldsymbol{X}}_i \widetilde{\boldsymbol{X}}_i'\right)$, and $\boldsymbol{\Omega} = \mathrm{E}\left(\boldsymbol{V}_i \boldsymbol{V}_i'\right)$, then

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \boldsymbol{\beta} + \left(n^{-1}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left[n^{-1}\boldsymbol{X}'(\boldsymbol{U} - \boldsymbol{V}\boldsymbol{\beta})\right] \\ &\xrightarrow{p} \boldsymbol{\beta} - (\boldsymbol{Q} + \boldsymbol{\Omega})^{-1}\boldsymbol{\Omega}\boldsymbol{\beta} \neq \boldsymbol{\beta}, \end{aligned}$$

  under the assumptions that $\mathrm{E}\left(\widetilde{\boldsymbol{X}}_i \boldsymbol{V}_i'\right) = 0$, $\mathrm{E}\left(\widetilde{\boldsymbol{X}}_i u_i\right) = 0$, and $\mathrm{E}\left(\boldsymbol{V}_i u_i\right) = 0$.

Endogeneity Issue: Simultaneous Equation Bias

- Consider following supply and demand equations

$$
\begin{aligned}
Q_t^d &= \beta_0 + \beta_1 P_t + \varepsilon_t^d, \quad \beta_1 < 0 \\
Q_t^s &= \alpha_0 + \alpha_1 P_t + \varepsilon_t^s, \quad \alpha_1 > 0 \\
Q_t^d &= Q_t^s \quad (= Q_t)
\end{aligned}
$$

  and $\varepsilon_t^d \overset{\text{i.i.d.}}{\sim} (0, \sigma_d^2)$, $\varepsilon_t^s \overset{\text{i.i.d.}}{\sim} (0, \sigma_s^2)$, $\text{Cov}\left(\varepsilon_t^d, \varepsilon_t^s\right) = 0$.

- Suppose you observe $\{P_t\}_{t=1}^T$ and $\{Q_t\}_{t=1}^T$ and run following regression

$$Q_t = \delta_0 + \delta_1 P_t + u_t.$$

Endogeneity Issue: Simultaneous Equation Bias

- We cannot either pin down β_1 or α_1 using OLS estimator and $\hat{\delta}_1$.
- First, we solve Q_t and P_t by representing the simultaneous equations as follows

$$\begin{bmatrix} 1 & -\beta_1 \\ 1 & -\alpha_1 \end{bmatrix} \begin{bmatrix} Q_t \\ P_t \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \alpha_0 \end{bmatrix} + \begin{bmatrix} \varepsilon_t^d \\ \varepsilon_t^s \end{bmatrix}$$

so that

$$\begin{bmatrix} Q_t \\ P_t \end{bmatrix} = \begin{bmatrix} 1 & -\beta_1 \\ 1 & -\alpha_1 \end{bmatrix}^{-1} \left[\begin{pmatrix} \beta_0 \\ \alpha_0 \end{pmatrix} + \begin{pmatrix} \varepsilon_t^d \\ \varepsilon_t^s \end{pmatrix} \right].$$

Endogeneity Issue: Simultaneous Equation Bias

- Next, note that

$$\begin{bmatrix} \hat{\delta}_0 \\ \hat{\delta}_1 \end{bmatrix} = \begin{bmatrix} T & \sum_{t=1}^{T} P_t \\ \sum_{t=1}^{T} P_t & \sum_{t=1}^{T} P_t^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{t=1}^{T} P_t \\ \sum_{t=1}^{T} P_t Q_t \end{bmatrix}$$

  hence,

$$\hat{\delta}_1 = \frac{-\left(\sum_{t=1}^{T} P_t\right)^2 + T \sum_{t=1}^{T} P_t Q_t}{T \sum_{t=1}^{T} P_t^2 - \left(\sum_{t=1}^{T} P_t\right)^2}$$

- We can show that (Exercise)

$$\hat{\delta}_1 \xrightarrow{p} \frac{(\alpha_0 - \beta_0)(\beta_1 \alpha_0 - \beta_0 \alpha_1)}{(\beta_1 - \alpha_1)^2 (\sigma_s^2 + \sigma_d^2)} + \frac{\sigma_s^2}{\sigma_s^2 + \sigma_d^2} \beta_1 + \frac{\sigma_d^2}{\sigma_s^2 + \sigma_d^2} \alpha_1.$$

Instrument Variable Estimation

- One major problem arsing from endogeneity is that coefficients of interested variables are not consistent.
- Suppose that X_j refers to the endogenous variable in linear regression model

$$Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + u_i,$$

an instrument variable Z is a random variable such that

(1) $\text{Cov}(Z, X_j) \neq 0$.
(2) $\text{Cov}(Z, u) = 0$.
(3) Z is not highly correlated with other variables $X_{\setminus j}$.

- The key idea of instrument variable estimation originates from moment estimation methods.

Instrument Variable Estimation

- For the simple univariate linear regression model, $\mathrm{E}(u_i) = 0$ and $\mathrm{Cov}(Z_i, u_i) = 0$, sample moment condition is

$$\frac{1}{n}\sum\left(Y_i - \tilde{\beta}_0 - \tilde{\beta}_1 X_i\right) = 0 \quad \frac{1}{n}\sum Z_i\left(Y_i - \tilde{\beta}_0 - \tilde{\beta}_1 X_i\right) = 0.$$

Hence the normal system of equations can be established

$$\begin{aligned}\sum Y_i &= n\tilde{\beta}_0 + \tilde{\beta}_1 \sum X_i \\ \sum Z_i Y_i &= \tilde{\beta}_0 \sum Z_i + \tilde{\beta}_1 \sum Z_i X_i\end{aligned}$$

and finally solves

$$\widetilde{\beta}_1 = \frac{\sum z_i y_i}{\sum z_i x_i}, \quad \widetilde{\beta}_0 = \bar{Y} - \widetilde{\beta}_1 \bar{X}.$$

Instrument Variable Estimation

- In general, suppose that we can partition

$$\boldsymbol{X}_i = (1, X_{i1}, X_{i2}, \cdots, X_{ik})'$$

into two parts such that

$$\boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{X}_{i1} \\ \boldsymbol{X}_{i2} \end{pmatrix} \begin{matrix} k_1 \times 1 \\ k_2 \times 1 \end{matrix}$$

where $\boldsymbol{X}_{i1}$ collects exogenous variables such that $\mathrm{E}(\boldsymbol{X}_{i1} u_i) = 0$ and $\boldsymbol{X}_{i2}$ collects endogenous variables such that $\mathrm{E}(\boldsymbol{X}_{i2} u_i) \neq 0$.

Instrument Variable Estimation

- A random vector $\boldsymbol{Z}_i$ $(l \times 1)$ is an instrument variable if
  (1) $\mathrm{E}(\boldsymbol{Z}_i u_i) = 0$
  (2) $\mathrm{E}(\boldsymbol{Z}_i \boldsymbol{X}_i') \neq 0$
- Given the definition of $\boldsymbol{Z}_i$, $\boldsymbol{X}_{i1}$ should be included in $\boldsymbol{Z}_i$. We can denote it as $\boldsymbol{Z}_{i1} \equiv \boldsymbol{X}_{i1}$, the **included exogenous variables**.
- We can define the remained part of $\boldsymbol{Z}_i$, with $\boldsymbol{Z}_{i1}$ excluded, as $\boldsymbol{Z}_{i2}$ so that for each i we have $\boldsymbol{Z}_i = (\boldsymbol{Z}_{i1}', \boldsymbol{Z}_{i2}')'$. $\boldsymbol{Z}_{i2}$ is referred to as the **excluded exogenous variables**.
- We define $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)'$, an $n \times l$ matrix.
- In matrix notation:

$$\underset{n\times k}{\boldsymbol{X}} = \underset{n\times k_1}{\left[\boldsymbol{X}^{(1)}}, \underset{n\times k_2}{\boldsymbol{X}^{(2)}}\right], \quad \underset{n\times l}{\boldsymbol{Z}} = \underset{n\times l_1}{\left[\boldsymbol{Z}^{(1)}}, \underset{n\times l_2}{\boldsymbol{Z}^{(2)}}\right]$$

Instrument Variable Estimation

- Recall that

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} \tag{†}$$

  and we can assume that

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Gamma} + \mathcal{E} \tag{‡}$$

  where $\mathcal{E} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)'$ is a $n \times k$ matrix collecting projection error term.
- $\mathrm{E}(\boldsymbol{Z}_i \boldsymbol{\varepsilon}_i') = 0$ given the definition of projection error.

Instrument Relevance Condition $(*)$

- From $(\ddagger)$, for a specific i we have
$$\boldsymbol{X}_i = \boldsymbol{\Gamma}' \boldsymbol{Z}_i + \boldsymbol{\varepsilon}_i.$$
- Since $\mathrm{E}\left(\boldsymbol{Z}_i \boldsymbol{\varepsilon}_i'\right) = 0$, we can derive that
$$\boldsymbol{\Gamma} = \mathrm{E}\left(\boldsymbol{Z}_i \boldsymbol{Z}_i'\right)^{-1} \mathrm{E}\left(\boldsymbol{Z}_i \boldsymbol{X}_i'\right).$$
and hence $\mathrm{E}\left(\boldsymbol{Z}_i \boldsymbol{X}_i'\right) \neq 0 \Rightarrow \boldsymbol{\Gamma} \neq 0$.
- If we further assume that $\mathrm{E}\left(\boldsymbol{\varepsilon}_i\right) = 0$, we can further derive that
$$\begin{aligned} \boldsymbol{\Gamma} \;=\; & \left\{\mathrm{E}\left[\boldsymbol{Z}_i - \mathrm{E}\left(\boldsymbol{Z}_i\right)\right]\left[\boldsymbol{Z}_i - \mathrm{E}\left(\boldsymbol{Z}_i\right)\right]'\right\}^{-1} \\ & \times \mathrm{E}\left[\boldsymbol{Z}_i - \mathrm{E}\left(\boldsymbol{Z}_i\right)\right]\left[\boldsymbol{X}_i - \mathrm{E}\left(\boldsymbol{X}_i\right)\right]'. \end{aligned}$$
and hence $\mathrm{Cov}\left(\boldsymbol{Z}_i, \boldsymbol{X}_i'\right) \neq 0 \Rightarrow \boldsymbol{\Gamma} \neq 0$.
- In this sense, $\mathrm{Cov}\left(\boldsymbol{Z}_i, \boldsymbol{X}_i'\right) \neq 0$ is a more restrictive instrument relevance condition.

Instrument Variable Estimation

- By substituting (‡) into (†), we can establish the relationship between $\boldsymbol{Y}$ and $\boldsymbol{Z}$

$$\begin{aligned} \boldsymbol{Y} &= (\boldsymbol{Z}\boldsymbol{\Gamma} + \mathcal{E})\,\boldsymbol{\beta} + \boldsymbol{u} \\ &= \boldsymbol{Z}\boldsymbol{\Gamma}\boldsymbol{\beta} + \mathcal{E}\boldsymbol{\beta} + \boldsymbol{u} \\ &= \boldsymbol{Z}\boldsymbol{\lambda} + \mathbf{V} \qquad (*) \end{aligned}$$

- OLS for (‡) yields $\hat{\boldsymbol{\Gamma}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\,\boldsymbol{Z}'\boldsymbol{X}$.
- OLS for $(*)$ yields $\hat{\boldsymbol{\lambda}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\,\boldsymbol{Z}'\boldsymbol{Y}$.
- It would be natural to ask whether we can solve $\boldsymbol{\beta}$ from $\boldsymbol{\Gamma}\boldsymbol{\beta} = \boldsymbol{\lambda}$ as an estimator for $\boldsymbol{\beta}$.

Instrument Variable Estimation: Identification

- If one can solve or recover $\mathbf{\Gamma}\boldsymbol{\beta} = \boldsymbol{\lambda}$, we say that $\boldsymbol{\beta}$ is **identified**.
- Assume that $\text{rank}\,(\mathbf{\Gamma}) = k$, then
  - if $l = k$, $\boldsymbol{\beta} = \mathbf{\Gamma}^{-1}\boldsymbol{\lambda}$, which refers to the **just-identified** case.
  - if $l > k$, then for any positive definite matrix $\boldsymbol{W}$,

$$\boldsymbol{\beta} = \left(\mathbf{\Gamma}'\boldsymbol{W}\mathbf{\Gamma}\right)^{-1}\mathbf{\Gamma}'\boldsymbol{W}\boldsymbol{\lambda},$$

    which refers to the **over-identified** case.
- A necessary condition (not sufficient) for $\text{rank}\,(\mathbf{\Gamma}) = k$ is $l \geq k$.

Instrument Variable Estimation: Identification

- For the just-identified case, we have

$$\hat{\boldsymbol{\beta}}_{IV} = (\boldsymbol{Z}'\boldsymbol{X})^{-1}\boldsymbol{Z}'\boldsymbol{Y}.$$

- $\hat{\boldsymbol{\beta}}_{IV}$ is consistent for $\boldsymbol{\beta}$ under the regular conditions.
- Finding a good IV(s) is more like **Art** rather than **Science**. For instance, Qian (2008, QJE) *"Missing Women and the Price of Tea in China: The Effect of Sex-Specific Earnings on Sex Imbalance"*

Two Stage Least Squares (2SLS)

- **Stage 1**: Regress $\boldsymbol{X}$ on $\boldsymbol{Z}$ to obtain $\hat{\boldsymbol{\Gamma}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{X}$ and save the predicted value $\hat{\boldsymbol{X}} = \boldsymbol{Z}\hat{\boldsymbol{\Gamma}} = P_{\boldsymbol{Z}}\boldsymbol{X}$.
- **Stage 2**: Regress $\boldsymbol{Y}$ on $\hat{\boldsymbol{X}}$ to obtain the 2SLS estimator $\hat{\boldsymbol{\beta}}_{2SLS} = \left(\hat{\boldsymbol{X}}'\hat{\boldsymbol{X}}\right)^{-1}\hat{\boldsymbol{X}}'\boldsymbol{Y}$.
- In fact,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{2SLS} &= (\boldsymbol{X}'P_{\boldsymbol{Z}}\boldsymbol{X})^{-1}\boldsymbol{X}'P_{\boldsymbol{Z}}\boldsymbol{Y} \\ &= \left[\frac{\boldsymbol{X}'\boldsymbol{Z}}{n}\left(\frac{\boldsymbol{Z}'\boldsymbol{Z}}{n}\right)^{-1}\frac{\boldsymbol{Z}'\boldsymbol{X}}{n}\right]^{-1}\frac{\boldsymbol{X}'\boldsymbol{Z}}{n}\left(\frac{\boldsymbol{Z}'\boldsymbol{Z}}{n}\right)^{-1}\frac{\boldsymbol{Z}'\boldsymbol{Y}}{n}.\end{aligned}$$

- Under the regular conditions, $\hat{\boldsymbol{\beta}}_{2SLS}$ is consistent for $\boldsymbol{\beta}$.

For the just-identified case, i.e. $l = k$

- $\hat{\boldsymbol{\beta}}_{2SLS} = \hat{\boldsymbol{\beta}}_{IV}$.

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{2SLS} &= \left[\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right]^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{Y} \\
&= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)\left(\boldsymbol{X}'\boldsymbol{Z}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{Y} \\
&= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{Z}'\boldsymbol{Y}
\end{aligned}$$

- We can solve IV estimator as $\hat{\boldsymbol{\beta}}_{IV} = \hat{\boldsymbol{\Gamma}}^{-1}\hat{\boldsymbol{\lambda}}$.

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{ILS} &= \hat{\boldsymbol{\Gamma}}^{-1}\hat{\boldsymbol{\lambda}} \\
&= \left[\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{X}\right)\right]^{-1}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{Y}\right) \\
&= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{Y}\right) \\
&= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{Z}'\boldsymbol{Y} = \hat{\boldsymbol{\beta}}_{IV}
\end{aligned}$$

Qian (2008, QJE)

Background:

- How economic conditions affect sex imbalance?
- **Identification problems:** areas with higher female income may have higher income precisely because women's status is higher for other reasons.
- How to find appropriate proxy variables?

Qian (2008, QJE)

- Two reforms after 1979: (1) increase in procurement prices of cash crops; (2) Household Production Responsbility System (HPRS).
- Women have a comparative advantage in producing tea, whereas men have a comparative advantage in producing orchard fruits. Therefore, areas suitable for tea cultivation experienced an increase in female-generated income, whereas areas suitable for orchard cultivation experienced an increase in male-generated income.

Qian (2008, QJE)

Key regression function:

$$sex_{ic} = (tea_i \times post_c)\,\beta + (orchard_i \times post_c)\,\delta + (cashcrop_i \times post_c)\,\rho + Han_{ic}\zeta + \alpha + \psi_i + \gamma_c + \varepsilon_{ic}$$

- sex_{ic}: fraction of males in country i of cohort c (individuals born after 1979 or not, $post_c$ as the dummy variable).
- tea_i: the amount of tea planted.
- $orchard_i$: the amount of orchard planted.
- $cashcrop_i$: the amount of cash crops planted.
- Han_i: the fraction of ethnically Han.

Qian (2008, QJE)

Two reasons for IV:

- 1997 agricultural data to proxy for agricultural conditions in earlier years introduces measurement error.
- Many other reasons, not necessarily only the increased value of female labor, that may lead to increase of amount of tea planted after 1979 reform. In other words, we potentially have omitted variable bias.

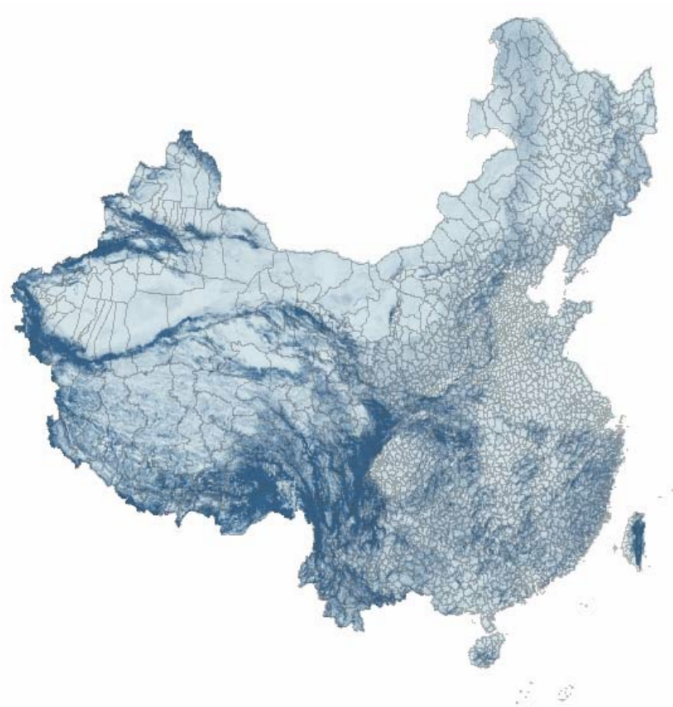
Inspiring IV in Qian (2008, QJE):

- Using the **average slope** of each county as the IV for tea planting.

Qian (2008, QJE)

Darker shades correspond to more tea planted per household.

Darker shades correspond to steeper regions.



Qian (2008, QJE)

TABLE III
OLS AND 2SLS ESTIMATES OF THE EFFECT OF PLANTING TEA AND ORCHARDS ON SEX RATIOS CONTROLLING FOR COUNTY LEVEL LINEAR COHORT TRENDS

	Dependent variables					
	Fraction of males			Tea × post	Fraction of males	
	(1) OLS	(2) OLS	(3) OLS	(4) 1st	(5) IV	(6) IV
Tea × post	−0.012	−0.013	−0.012		−0.072	−0.011
	(0.007)	(0.006)	(0.005)		(0.031)	(0.007)
Orchard × post	0.005					
	(0.002)					
Slope × post	−0.002			0.26		
	(0.002)			(0.057)		
Linear trend	No	No	Yes	Yes	No	Yes
Observations	28,349	37,756	37,756	37,756	37,756	37,756

Test of Endogeneity

Durbin-Wu-Hausman tests:

- Consider testing:

$$\begin{aligned} H_0 &: \mathrm{E}\left(\boldsymbol{X}_i u_i\right) = 0 \\ H_1 &: \mathrm{E}\left(\boldsymbol{X}_i u_i\right) \neq 0, \quad \mathrm{E}\left(\boldsymbol{Z}_i u_i\right) = 0 \end{aligned}$$

- Recall that

$$\underset{n\times k}{\boldsymbol{X}} = \Big[\underset{n\times k_1}{\boldsymbol{X}^{(1)}}, \underset{n\times k_2}{\boldsymbol{X}^{(2)}}\Big], \quad \underset{n\times l}{\boldsymbol{Z}} = \Big[\underset{n\times l_1}{\boldsymbol{Z}^{(1)}}, \underset{n\times l_2}{\boldsymbol{Z}^{(2)}}\Big]$$

and run following **augmented** regression

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + P_{\boldsymbol{Z}}\boldsymbol{X}^{(2)}\boldsymbol{\delta} + \boldsymbol{\varepsilon}.$$

Test of Endogeneity

- **Step 1**. Run the OLS regression of $\boldsymbol{X}^{(2)}$ on the instrumental variables $\boldsymbol{Z}$ and obtain the fitted value $\widehat{\boldsymbol{X}}^{(2)} = P_{\boldsymbol{Z}}\boldsymbol{X}^{(2)}$.
- **Step 2**. Run the OLS regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$ and $\widehat{\boldsymbol{X}}^{(2)}$ and test whether the coefficient vector $\boldsymbol{\delta}$ of $\hat{\boldsymbol{X}}^{(2)}$ is 0.
- Under H_0, the F-test statistic F_n based on the **augmented** regression is asymptotically distributed as the Wald statistic: $W_n \equiv k_2 F_n$ is asymptotically distributed as $\chi^2(k_2)$.