Income and Geographic Segregation Across Colleges in China *

Mingyu Chen† Ernest Liu‡ Yinshan Shang§

November 30, 2021

[Preliminary draft, click here for the latest version.]

Abstract

College access has been a central topic in economics and education as higher education is a key driver of social mobility, especially for low-income students. We study segregations in the Chinese college admission system, where the distribution of seats depends on centralized exams and location quotas. We construct a novel county-level income dataset covering nearly all counties in China and link it to the universe of college-admitted students in 2006-2011. We find that the admission rate to general college education is equal across all income groups. While students from low-income counties are disadvantaged in going to a handful of very top universities with a strong preference towards local students, the disadvantage nearly disappears when elite colleges are more generally defined (top 100 schools admitting less than 8% of students). We build a structural model that exploits geographic variations of income and test performances to estimate income inequality within counties and find similar results when using county-level data directly.

*Any errors are our own.

†Chen: Princeton University, mingyuc@princeton.edu.
‡Liu: Princeton University, ernestliu@princeton.edu.
§Shang: Princeton University, yinshans@princeton.edu.

1 Introduction

College education is widely believed to be an important device for upward mobility. Income segregation in college admission is concerning not just for moral reasons; difficulty in accessing higher education prevents low-income students from realizing their potential productivity and leads to inefficiency for society.

Moreover, when low-income students attend colleges, do they have equal access to elite colleges as their wealthy peers? Graduates of elite colleges attain higher earnings and occupy a large share of the managerial positions (Li et al., 2012; Zimmerman, 2019). If colleges with the best earnings outcome mainly admit wealthy students, the college education system could aggravate the persistence of income inequality across generations instead of attenuating it.

In this paper, we document the nationwide income and geographic segregation across colleges in China. We focus on low-income students' access to highly-ranked universities, which is an important pathway to upward mobility. To do this, we match all students admitted to colleges in 2006-2011 to the county's average income they are from, defined by the place they registered for the National College Entrance Exam (NCEE). The county average income serves as a proxy of a student's background. We then measure the admission rate to different tiers of universities for all income groups. To incorporate the within county inequalities, we exploit geographic variations of income and test performance across counties and update estimates of admission inequality based on the correlation between income and performance.

We find that the overall college admission rate is equal across all income groups. The admission rate to four-year colleges is slightly increasing in income rank, while the admission rate to three-year colleges is slightly decreasing in income. However, the magnitude is relatively small. When zooming in to different college tiers, students from low-income counties are disadvantaged in going to the several top colleges, but the disadvantage nearly disappears if we extend coverage to around top 100 elite colleges[1]. For the several top colleges, the income segregation is largely driven by preference towards local students. The updated estimates of inequality are very close to that using county-level data directly, implying that most inequality arises from differences across geographic regions instead of across income groups within the same region.

Our study is related to two strands of literature. The first is studies on income segre-

[1]Here we include the colleges in Project 211, around top 100 colleges in China. For the rest of the paper, we refer to Project 211 colleges as elite colleges.

gation across colleges. Chetty et al. (2020) found strong segregation in parental income across colleges in the US; colleges that produce high-income graduates tend to admit disproportionately students from affluent families. We find less segregation in the Chinese admission data. Top colleges in China also tend to admit more wealthy kids, but the difference is smaller than that found by Chetty et al. (2020). Li et al. (2015) estimated the income segregation in college admission in China using data from 2003. Consistent with our result, they concluded that income segregation is not large; most of the segregation they identified arise due to urban-rural differences. Hoxby and Avery (2012) focused on how lack of information leads low-income, high-achieving students in the US to choose lower-ranked colleges than they could have attended. We aim to decompose the observed inequality in Chinese data and measure the impact of information asymmetry in a system with more transparent admission criteria. Chetty et al. (2016) found that moving to a low-poverty neighborhood increases students' college attendance rate. Using Chinese data, we find nearly uniform college attendance rates across counties with different income levels.

The second is studies on the Chinese higher educations system and its mobility implications. Yang (2021) studied the impact of admission quota and found it mitigating geographic inequality across provinces at the expense of less efficiency in human capital output. Our study does not directly study the impact of quota policy; we depict the current state of inequality given the quota policies. Our finding suggests that access to college education is equal under the current quota allocation except for the several top colleges. Li et al. (2012) estimated returns to college education and found that the gross return to attending the top 100 elite colleges is 10.7%. Our study finds that admission to elite colleges is equal across income groups. Combined with Li et al. (2012), this suggests the upward mobility path through attending top colleges in China is accessible to low-income students.

We contribute to the literature in three aspects. First, our study is the first to measure income segregation in China's college admission system comprehensively. Combined with studies on college education returns, we also offer implications to the current state of intergenerational upward mobility through higher education. The Chinese college admission system is unique in its centralized exam and geographic quota policies. Students are purely admitted based on NCEE scores, so the selection process is income-neutral except for the underlying correlation between parental income and exam performance. Chetty et al. (2020) concluded based on a simulation that if US colleges admit students purely based on the ACT/SAT scores, income segregation will fall significantly. Our study provides

an empirical benchmark with an income-neutral admission system. Due to difficulty in obtaining data, the income segregation across colleges in China is largely understudied. We devoted extensive effort and compiled a county-level income and population panel covering over 95% of any year in the college admission dataset.

Second, our study also offers insights into the impact of the college admission system on geographic mobility. China's unique household registration system (Hukou system) makes it difficult for workers to move and work at a new place. The college admission system is an important geographic mobility device as it allows students to migrate to another city to study and work after graduation. We analyze the composition of college students from local and non-local areas to understand the migration flow due to college attendance. We find that access to top colleges is largely equal across income levels for students not from the same province as the colleges. Top colleges are offering accessible opportunities for students from all income backgrounds to migrate.

Lastly, we provide a novel method to account for individual-level differences. We overcome the lack of individual-level income data by estimating a structural model with cross-region variations. In this study of admission inequality, this method generates results similar to that using county-level data directly, implying cross-region variation is the main source of observed inequality.

The rest of the paper is structured as follows: in section 2, we discuss the data sources used and cleaning method; in section 3, we detail the estimation method used to incorporate within-county inequality; in section 4, we summarize findings; finally we conclude in section 5. Additional results and robustness checks can be found in the online appendix.

2 Data

2.1 Data sources

Our college admission data covers all students admitted to four-year colleges in China through the NCEE from 2005 to 2007, and all students admitted to any college (three-year or four-year) from 2008 to 2011. Main variables in the college admission dataset includes students' scores in the NCEE, their test types (either science or humanities), the college and major they are admitted to, and type of the college (either three-year or four-year). For 2009 to 2011, we miss data from specific college types for a few provinces. Detailed missing data notes can be found in Appendix A. The main results are robust to using a balanced panel and excluding provinces with missing data, see Appendix for robustness

checks.

We construct a panel of county level GDP per capita and residential population. To construct the county level panel, we combine data from China Statistical Yearbook for Regional Economy and China County Statistical Yearbook (collected by the EPS China Regional Economy database), Municipal and National Bureau of Statistics (collected by the CEIC database), provincial statistical yearbooks, WIND database, and county level Statistical Communiqué. Detailed coverage information and comparison of the data sources can be found in Appendix A.

We combine data for college admitted students from 2005-2011 with county level income and population. We identify the county each student is from through extracting the county codes in their test-taker number, corresponding to the place the student took the NCEE. County code is a six-digit ID assigned by the Chinese Ministry of Civil Affairs to each county-level administrative region. We then match the six-digit county codes to county names with crosswalks from the Ministry of Civil Affairs, and join the college admission dataset with county level income and population data using county names.

2.2 Summary Statistics

Universities in China can be broadly categorized into four-year and three-year colleges. Four year colleges are generally academic-oriented universities, while three-year colleges are similar to community colleges in the US and focus on technical training. Table 1 shows the number of observations in the college admission dataset, as well as the coverage of county level panel. Each observation in the college admission dataset represents a student admitted to college. For all years and college types, the coverages of county level data are over 95%. More detailed information can be found in Appendix A.

Within all four-year colleges, nine top universities formed the C9 League, usually considered the Chinese equivalent of the US Ivy-League universities. The next tier is around 116 universities selected into Project 211 (meaning the top 100 universities for the 21st Century) initiated in 1995 by the Ministry of Education of the People's Republic of China. All C9 colleges are in Project 211; all colleges in Project 211 are four-year colleges. In the following text, we refer to the C9 League colleges as "elite colleges" and the colleges in Project 211 as "top colleges". Figure 1 shows the share and number of students admitted to the C9 League, the other top colleges, the other four-year colleges, and the three-year colleges. In Figure 1a, we omit years 2005 to 2007, since we do not have data on students admitted to three-year colleges.

3 Method for Individual Income Construction

In this section, we propose a method to estimate the inequality in college admissions and overcome the lack of individual level data. As mentioned in section 2.1, we do not have students' individual family income. We approximate family background with county level GDP per capita. This method gives a lower bound for the inequality in college admission because it ignores within-county inequality; the observed inequality in admission only captures the effect that better colleges admit more students from the rich counties than the poor counties. In reality, we would also expect that family income is positively correlated with NCEE performance in each county. Although colleges admit students solely based on score ranks, top schools is more likely to admit the rich kids in each county because they tend to perform better in the NCEE.

To incorporate the within-county inequality and provide a better estimate of students' background, we build a structural model about income distribution and how family income affect a student's test score. We then exploit variations in county level GDP per capita to estimate this model. Finally, we use scores to form a posterior belief about each student's family income, then aggregate this prediction to obtain the admitted students' profile by college.

See Figure 4a and Figure 4b for a comparison of the income distribution at college levels based on GDP per capita data alone (left) and updated belief using scores (right). Overall, the estimated admission inequality after updating our belief of students' background with their scores is close to using county level data directly.

3.1 Model

The underlying logic of this model is that scores (s) and family income (x) are correlated. We establish a model of how family income, together with other group characteristics Z (county, year, and test subject [2]) affect scores, estimate the model with county level data, and use this model to predict individual income.

Written in the form of Bayes' formula as Equation 3.1, our method is to calculate the left hand side, income distribution given score and group characteristics, by estimating the right hand side terms.

[2]In most provinces in China, students choose between arts and science and take the respective set of tests.

$$\mathbb{P}(x|s,Z) = \frac{\mathbb{P}(s|x,Z) \times \mathbb{P}(x|Z)}{\mathbb{P}(s|Z)} \tag{3.1}$$

$\mathbb{P}(s|Z)$ denotes the score distribution for each county, year, and test subject. $\mathbb{P}(x|Z)$ is the family income distribution for each county-year[3]. We assume that the income for each county and year follows the Log-normal distribution. Using the county level GDP per capita as first moment, and taking the Gini coefficient estimate from literature (Chen et al., 2010), we could pin down the income distribution for each county and year. $\mathbb{P}(s|x,Z)$ denotes the conditional distribution of scores s given family income x and other characteristics Z. This term will be estimated with a model of relationship between scores and income; coefficient values will be estimated using cross-county variation.

3.1.1 Income distribution at county level

Considering model simplicity and empirical goodness-of-fit, we assume income in each county-year to be distributed according to a log-normal distribution.[4]. We estimate the two parameters with 1) the average income in county, using the GDP per capita data, and 2) the inequality level in county, using Gini coefficients based on estimates from literature.

With Lognormal(μ, σ) distribution, the correspondence between Gini coefficient and parameter is

$$\text{Gini Coefficient} = erf(\frac{\sigma}{2}) \tag{3.2}$$

where erf is the Gauss Error Function

$$erf(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt \tag{3.3}$$

We can use them to calculate the value of σ using the Gini coefficient for China from

[3]We omit test subject here because we assume that family income is independent of the test subject the student chooses.

[4]The most commonly used two-parameter distributions are Pareto distribution, Log-normal distribution, and Gamma distribution. Pareto distribution has a good fit for the upper tail, but doesn't fit the whole range well; Log-normal distribution fits well for the middle-income range, but doesn't fit the tails well; Gamma distribution is similar to log-normal but fits the tails slightly better. (Majumder and Chakravarty, 1990) As we will discuss below, log-normal income assumption simplifies the theoretical formulas, so we assume income to be log-normally distributed.

literature. Next, we estimate the mean parameter μ based on

$$E[x] = \exp(\mu + \frac{\sigma^2}{2}) = \text{ GDP per capita in county.} \tag{3.4}$$

This gives us the estimated income distribution in each county; we can then aggregate the estimated distributions and obtain the income distribution for entire nation.

3.1.2 Score distribution given income

To make scores for different provinces and years comparable, we first standardize the tests scores based on the full score. Specifically, we calculate the standardized score s as

$$s = \text{logit}(\frac{\text{score}}{\text{full score}}) \tag{3.5}$$

$$= \log(\frac{\text{score}}{\text{full score} - \text{score}}) \in \mathbb{R} \tag{3.6}$$

We assume that for each given income level x and characteristics Z, the standardized score s follows a Normal distribution with mean $\alpha \log(x) + k(Z)$ and standard error $\beta(Z)$:

$$s \sim \mathcal{N}(\alpha \log(x) + k(Z), (\beta(Z))^2) \tag{3.7}$$

The most important parameter here is α. It is constant for all counties in the same year, capturing the average effect of income on score. More precisely, 1 percent increase in family income on average increases the student's standardized score s by α. $k(Z)$ and $\beta(Z)$ are assumed to differ across counties to capture county level heterogeneity. A higher $k(Z)$ means it is easier to score high in county Z; a high $\beta(Z)$ means there are more uncertainty in scores in county Z.

3.1.3 Score distribution within county

To estimate the above parameters, we need to compute the marginal distribution of coun To estimate α for each year and $k(Z), \beta(Z)$ for each county-year, we compound the distribution of score given income and characteristics $(s|x, Z)$ with the county level income distribution $(x|Z)$ to obtain the marginal distribution of standardized score s for each county $(s|Z)$.

Notice that

$$\begin{cases} s|x, Z & \sim \mathcal{N}(\alpha \log(x) + k(Z), (\beta(Z))^2) \\ log(x)|Z & \sim \mathcal{N}(\mu(Z), \sigma^2) \end{cases} \tag{3.8}$$

implies

$$s|Z \sim \mathcal{N}(\alpha\mu(Z) + k(Z), \alpha^2\sigma^2 + \beta^2(Z)). \tag{3.9}$$

We can thus estimate the parameters by Method of Moments, which we will revisit and detail in due course.

3.1.4 Posterior Belief of Individual Family Income Given Score

The final step is to update our belief of individual student's family income based on the student's test score and other observable characteristics (county, year, test subject). More specifically, we are interested in estimating the probability that income is less than or equal to x^*, given standardized score $s = s^*$ and other characteristics Z:

$$\mathbb{P}(x \leq x^* | s = s*, Z). \tag{3.10}$$

To do this, we calculate the conditional density of log income (lx) given scores and other characteristics $g(lx = lx^*|s^*, Z)$.[5] Let g denote the probability density of an event, then

$$g(lx = lx^*|s^*, Z) = \frac{g(lx = lx^*, s = s^*|Z) \times g(lx = lx^*|Z)}{g(s = s^*|Z)} \tag{3.11}$$

$$= \frac{\frac{1}{\beta_z\sqrt{2\pi}} \exp(-\frac{1}{2}(\frac{s^* - \alpha lx^* - k_z}{\beta_z})^2) \times \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2}(\frac{lx^* - \mu_z}{\sigma})^2)}{\frac{1}{\sqrt{2\pi(\alpha^2\sigma^2 + \beta_z^2)}} \exp(-\frac{1}{2}(\frac{lx^* - k_z - \alpha\mu_z}{\sqrt{\alpha^2\sigma^2 + \beta_z^2}})^2)} \tag{3.12}$$

$$= \frac{\sqrt{\alpha^2\sigma^2 + \beta_z^2}}{\beta_z\sigma\sqrt{2\pi}} \exp(-\frac{1}{2}(\frac{lx^* - \frac{\alpha(s^* - k_z)^2\sigma^2 + \beta_z^2\mu_z}{\alpha^2\sigma^2 + \beta_z^2}}{\frac{\beta_z\sigma}{\sqrt{\alpha^2\sigma^2 + \beta_z^2}}})^2) \tag{3.13}$$

[5] We use log income instead of income because log income follows a normal distribution and is easier to represent; the results will be the same since we are actually interested in the cumulative density rather than probability density.

Notice that the posterior belief of log income is a normal distribution:

$$(lx|s^*, Z) \sim \mathcal{N}(\frac{\alpha(s^* - k_z)^2\sigma^2 + \beta_z^2\mu_z}{\alpha^2\sigma^2 + \beta_z^2}, \frac{\beta_z\sigma}{\sqrt{\alpha^2\sigma^2 + \beta_z^2}}). \tag{3.14}$$

Based on this result, we can calculate the probability of any individual being in any income rank group, and aggregate the probability by college to obtain the posterior belief of student background at a given college.

3.2 Estimation

We estimate the model in the following steps:

1. Estimate the parameters for income distribution in each county-year:

   We first use Equation 3.2 and Equation 3.3 to calculate a uniform σ for all counties and years. Based on Chen et al. (2010), in 2005, the overall Gini coefficient of income in China is around 0.45; this corresponds to $\sigma = 0.27$. As detailed in Appendix **??**, the estimation is robust to alternative Gini coefficient assumptions.

   Next, we estimate the mean of log income μ for each county-year using Equation 3.4. This gives us full information about the income distribution at any county or year. Aggregating all counties (weighted by their population size) for a given year, we can get the income distribution for the nation at that year.

2. Estimate the parameters for score distribution given income using method of moments.

   (a) We first estimate the correlation between mean score and income α. We estimate α by estimating the following model:

   $$\bar{s_z} = \alpha\mu_z + \gamma_p + \epsilon_z \tag{3.15}$$

   where s_z is the mean score of students for a fixed county, year, and test subject z; μ_z is the estimated mean of income for a fixed county, year, and test subject z; γ_p stands for province fixed effect; ϵ is noise.

   This estimation gives us $\alpha = 0.097$. This means doubling the family income (increase by 100%) is associated with the standardized score s increasing by around 0.1. Figure 2b demonstrates the relation with scatter plot.

(b) We then use the alpha to estimate the county-year specific β_z and k_z using method of moments again based on Equation 3.9:

$$k_z = \mathbb{E}[s|z] - \alpha * \mu_z \tag{3.16}$$

$$\beta_z^2 = \mathbb{E}[s^2|z] - (\mathbb{E}[s|z])^2 - \alpha^2\sigma^2 \tag{3.17}$$

3. Calculate posterior belief of individual family income:

   Given k_z, β_z, and individual standardized score s, the posterior belief of the family income of an individual with score s and characteristics z is a log-normal distribution. Rewriting Equation 3.14 to be the distribution of income x, we have:

$$(x|s^*, Z) \sim \text{Lognormal}(\frac{\alpha(s^* - k_z)^2\sigma^2 + \beta_z^2\mu_z}{\alpha^2\sigma^2 + \beta_z^2}, \frac{\beta_z\sigma}{\sqrt{\alpha^2\sigma^2 + \beta_z^2}}). \tag{3.18}$$

   We can thus calculate the probability that each individual's family income fall into any income rank category based on Equation 3.18. For example, the probability that a student with score s and characteristics z is in the bottom 20% income group is

$$p = F(\log(x) <= \log(\bar{x})) \tag{3.19}$$

$$= \Phi(\frac{\log(\bar{x}) - \text{mean}}{\text{sd}}) \tag{3.20}$$

   where Φ is the standard normal cumulative density function; $\bar{x}$ is the upper bound for bottom 20% income group in the national distribution as demonstrated in Step 1; mean and sd refers to the first and second parameters in distribution 3.18 respectively.

   In this step, we remove the county-year-test subject with less than 5 students, as the distribution estimated would not be very precise. For these students, we use the county level income distribution (which is a log-normal estimated in step 1) as our posterior belief for their family income.

4. Obtain aggregate inequality in admission:

   The last step is to collapse the individual posterior beliefs to college-years or college tier-years to compare the posterior belief of income distribution across colleges or college tiers. The results will be shown in next section.

4 Results

We find a flat college admission rate across income groups. Among all admitted students, the density shows an inverse-U shape: the highest admission rate comes from the group between 60% and 80% income percentile. This inverse-U shape is caused by 1) lower income students having low admission rate in general, and 2) highest income students having low admission rate to the lower-ranked colleges.

Students from high-income regions have significant advantage over low-income region students in being admitted to a highly-ranked college. However, the difference is only large if we focus on the several best colleges. The admission rate becomes more equal if we extend to the 985 colleges (around top 40), even more uniform if we extend to the Project 211 colleges (around top 100). We conclude that students from low-income regions have a decent chance of being admitted to a reasonably good college.

4.1 Admitted sample

We first produce the density of admitted students from each income percentile. Figure 3 plots the income distribution of all admitted students (first subfigure), and decomposes the admitted students into 4-year colleges (Benke) and 3-year colleges (Zhuanke).

Among all admitted students, the density shows an inverse-U shape: the highest density comes from the group between 60% and 80% income percentile. When we decompose the admitted students sample into 4-year colleges and 3-year colleges, we could observe that this inverse-U shape is caused by 1) lower income students unable to go to colleges in general, and 2) highest income students won't go to the lower-ranked colleges. We see that the admission rate of high income students is not lower than other income groups for Benke; but for Zhuanke, the admission rate is almost uniformly decreasing in income.

4.2 Student composition by income groups

Figure 4a and Figure 4b compares the student background distribution using county level data directly and after estimation. The left figures use county level GDP per capita as student's family income and provides a lower bound for inequality; the right figures are posterior belief estimated using the aforementioned method. The overall results are similar. Admission is generally increasing in income for higher-ranked colleges and decreasing for lower-ranked colleges.

4.3 Density of student background by college tiers

Figure 5 and Figure 6 plot the density of students family income in the nine C9-League Chinese colleges and the top 2 Chinese colleges respectively, using county level GDP and the aforementioned estimation method. As above, the estimated results is close to using county level GDP.

The purple lines represent students from a different province than the college they attend; the gap between green and purple lines represents the local students. For the C9-League and the top 2 colleges, the income distribution of non-local students is much more equal than the local students. Most low-income students are non-local students, admitted based on the quota assigned to other provinces, while local students are mostly from the wealthiest families in the nation. While the local students have strong advantage in admission, the quota assigned to all other provinces seem to be fairly equal even for these top colleges.

Figure 7a revisits the cumulative density by college tiers for Benke students using county level GDP per capita; Figure 7b shows the results estimated with the above method. As above, the estimated results is close to using county level GDP.

5 Conclusion and Discussion

We find overall college admission rate across all income groups is nearly uniform. Students from high income regions have advantage in going to four-year colleges, while students from low income regions are more likely admitted to a three-year college. High income regions have higher admission rate for the highest-ranked universities. However, the advantage is no longer salient if we extend the range of college to the top 100 nationally. Combined with the literature on college education returns, this implies that students from poor regions in China have decent chance to upward mobility through high quality college education.

The inequality in Chinese college admission provide important implications for college admission system design. The purely score-based admission criteria simplifies the information gathering process of students when choosing between universities. The location-based quota system also likely contributed to the relatively low inequality cross geographic regions.

The college admission system plays an important role in China's future economic transformation. As the growing cost of labor pushes down the growth rate of the economy,

supply of high skilled labor will be an important determinant of China's future growth. Since the period analyzed in this paper, many reforms have happened in the Chinese college admission system, mostly unifying and centralizing the exams across provinces. What this implies for this giant economy and the rest of the world is an open and exciting area for future research.

References

1. Chen, Jiandong, Dai Dai, Ming Pu, Wenxuan Hou, and Qiaobin Feng (2010). "The trend of the Gini coefficient of China". *Brooks World Poverty Institute Working Paper* 109.

2. Chetty, Raj, John N Friedman, Emmanuel Saez, Nicholas Turner, and Danny Yagan (2020). "Income segregation and intergenerational mobility across colleges in the United States". *The Quarterly Journal of Economics* 135.3, pp. 1567–1633.

3. Chetty, Raj, Nathaniel Hendren, and Lawrence F Katz (2016). "The effects of exposure to better neighborhoods on children: New evidence from the Moving to Opportunity experiment". *American Economic Review* 106.4, pp. 855–902.

4. Hoxby, Caroline M and Christopher Avery (2012). *The missing" one-offs": The hidden supply of high-achieving, low income students*. Tech. rep. National Bureau of Economic Research.

5. Li, Hongbin, Prashant Loyalka, Scott Rozelle, Binzhen Wu, and Jieyu Xie (2015). "Unequal access to college in China: How far have poor, rural students been left behind?" *The China Quarterly* 221, pp. 185–207.

6. Li, Hongbin, Lingsheng Meng, Xinzheng Shi, and Binzhen Wu (2012). "Does attending elite colleges pay in China?" *Journal of Comparative Economics* 40.1, pp. 78–88.

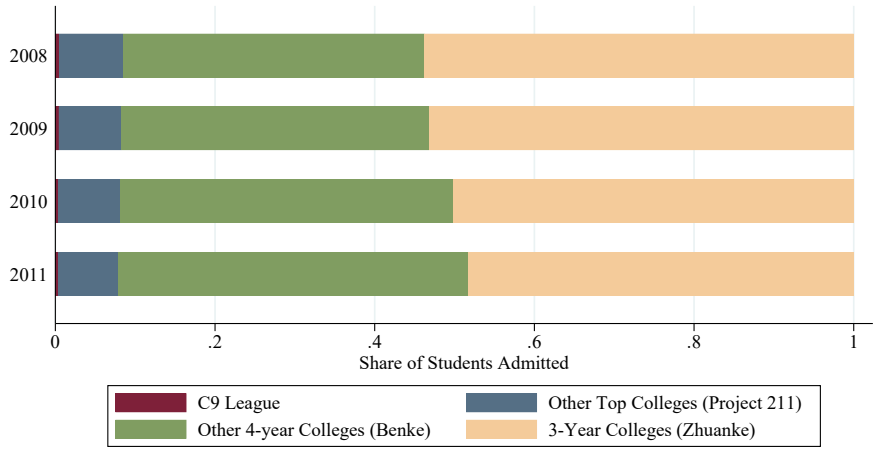
7. Majumder, Amita and Satya Ranjan Chakravarty (1990). "Distribution of personal income: Development of a new model and its application to US income data". *Journal of applied econometrics* 5.2, pp. 189–196.

8. Yang, Yu Alan (2021). "Place-Based College Admission, Migration and the Spatial Distribution of Human Capital: Evidence from China".

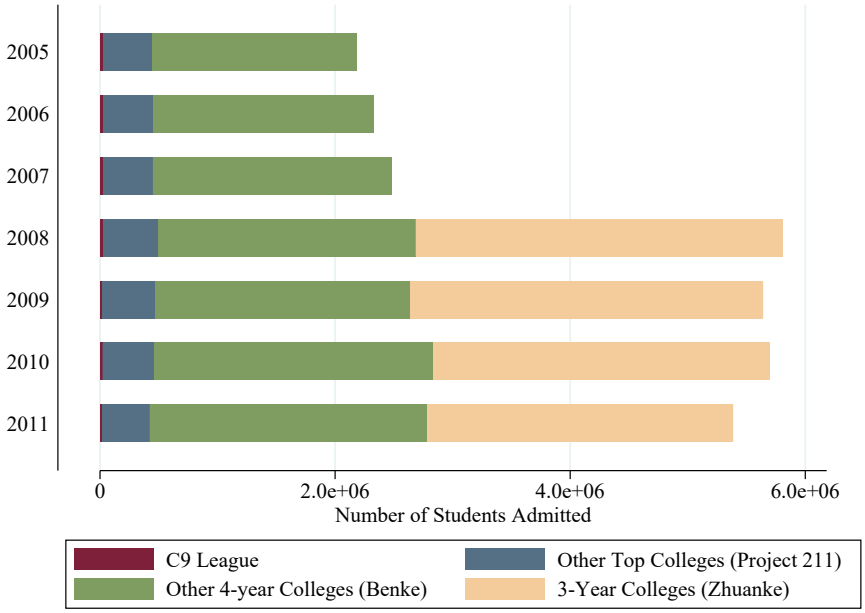
9. Zimmerman, Seth D (2019). "Elite colleges and upward mobility to top jobs and top incomes". *American Economic Review* 109.1, pp. 1–47.

Figures and tables

Figure 1: Composition of Chinese Universities

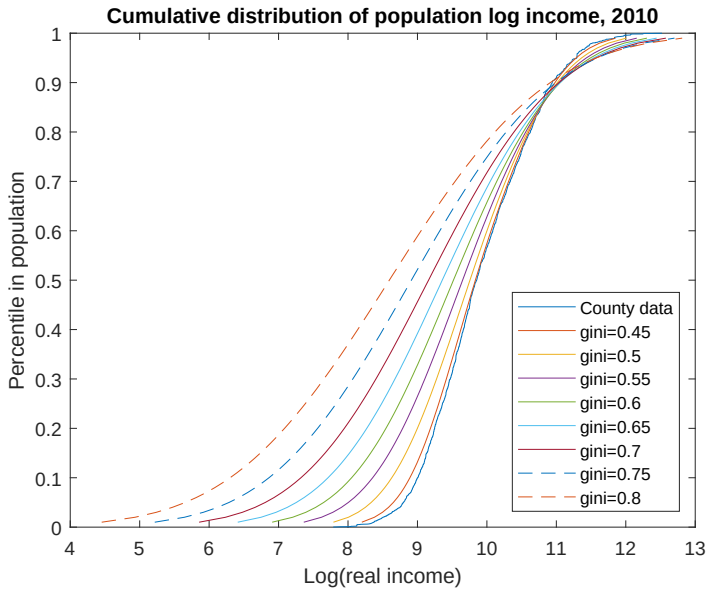


(a) Share of Students Admitted by College Tier and Year

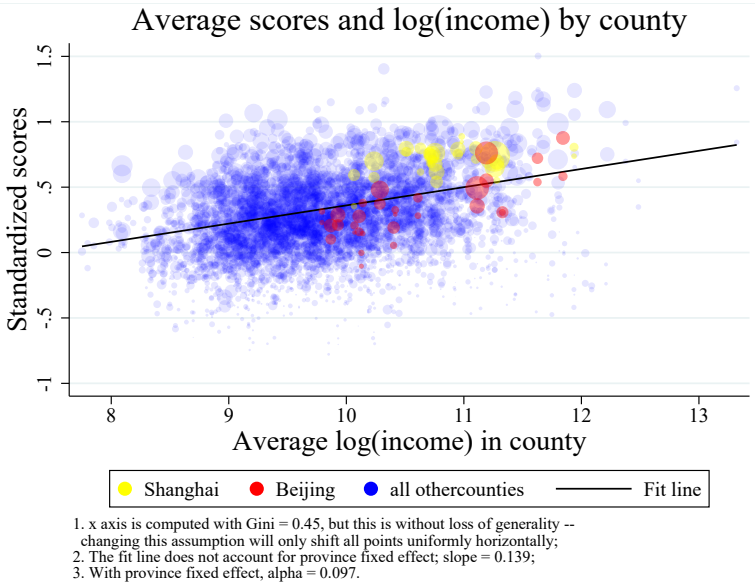


(b) Number of Students Admitted by College Tier and Year

Figure 2: Estimating within-County Segregation



(a) Estimated National Distribution of Log Income per Capita, 2010



(b) Average standardized score and county income, 2010

Figure 3: Density of Admitted Students' Family Income Rank, 2010

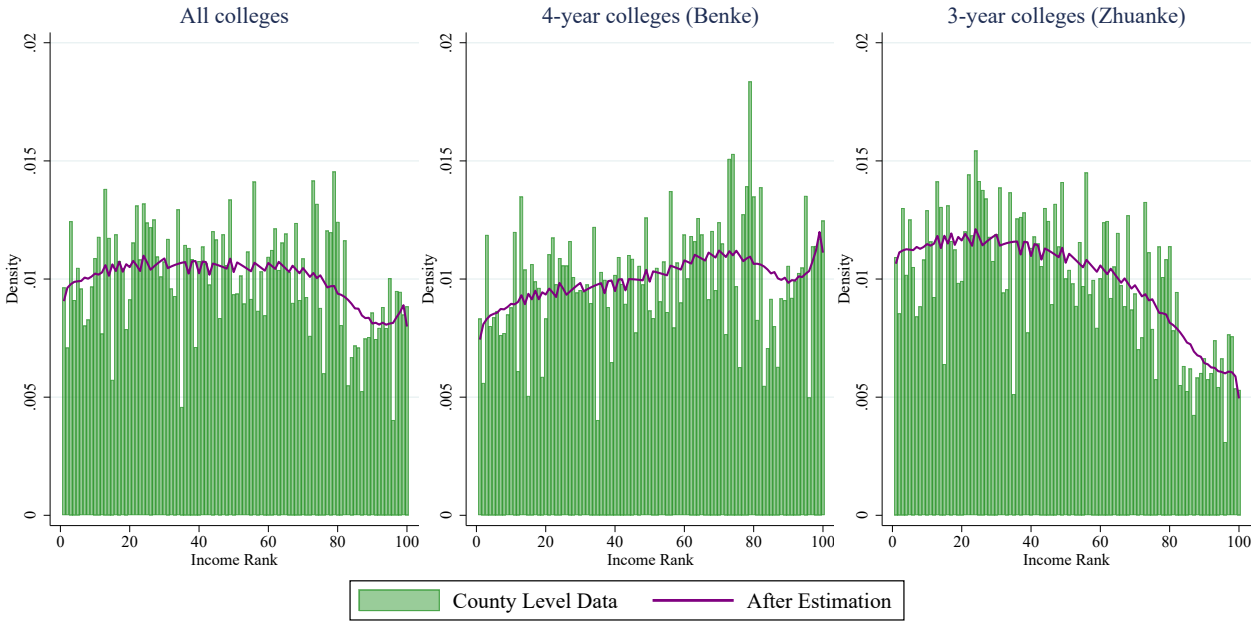
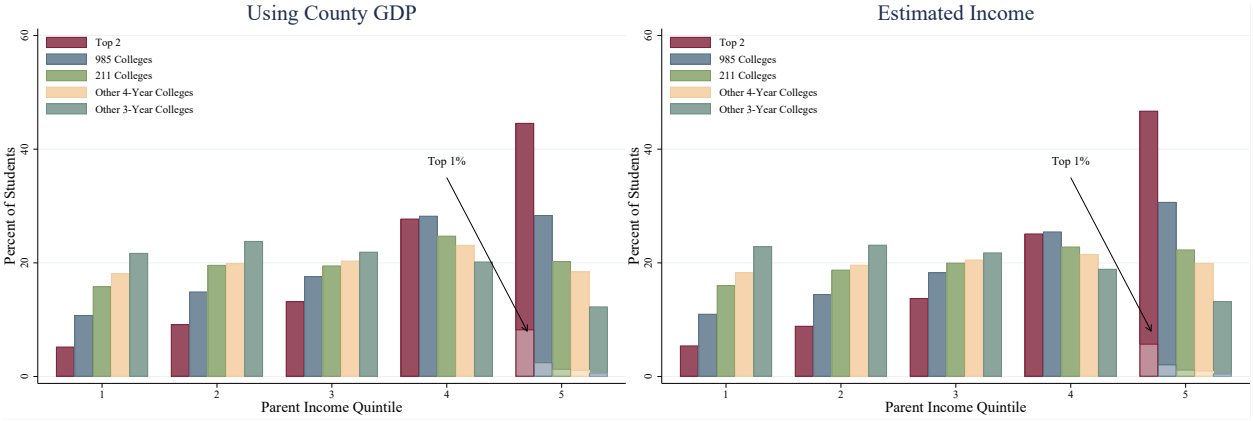


Figure 4: Student Background using County Level GDP and Estimation, 2010

(a) Five Groups Breakdown



(b) Ten Groups Breakdown

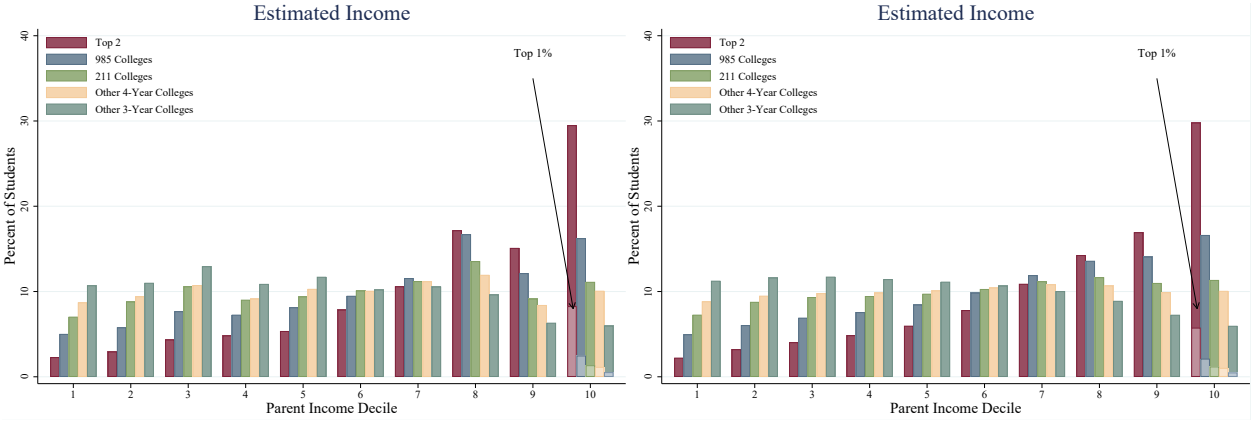


Figure 5: Density of Student Income at C9-League Colleges

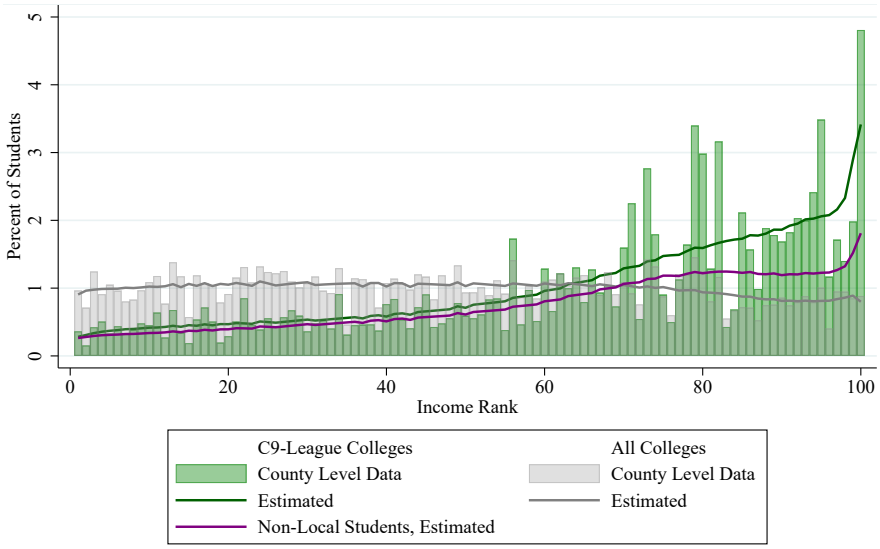


Figure 6: Density of Student Income at Top 2 Colleges

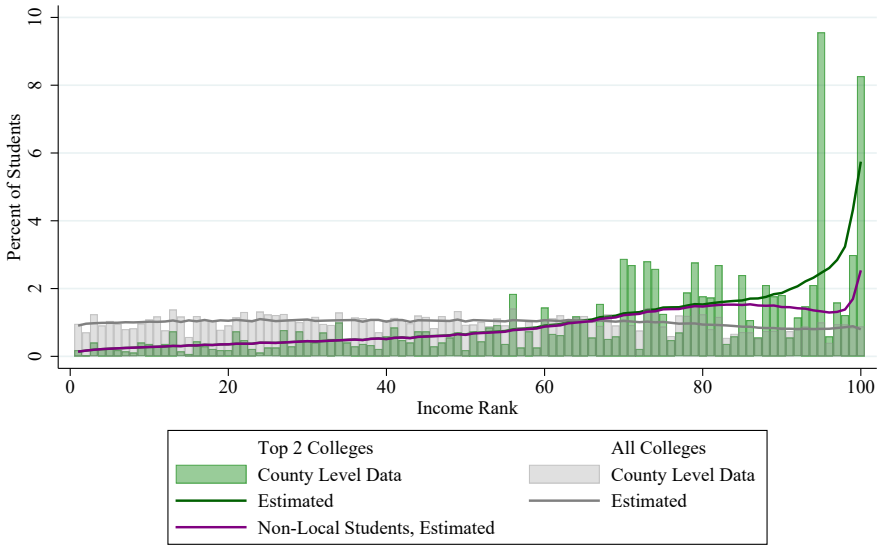
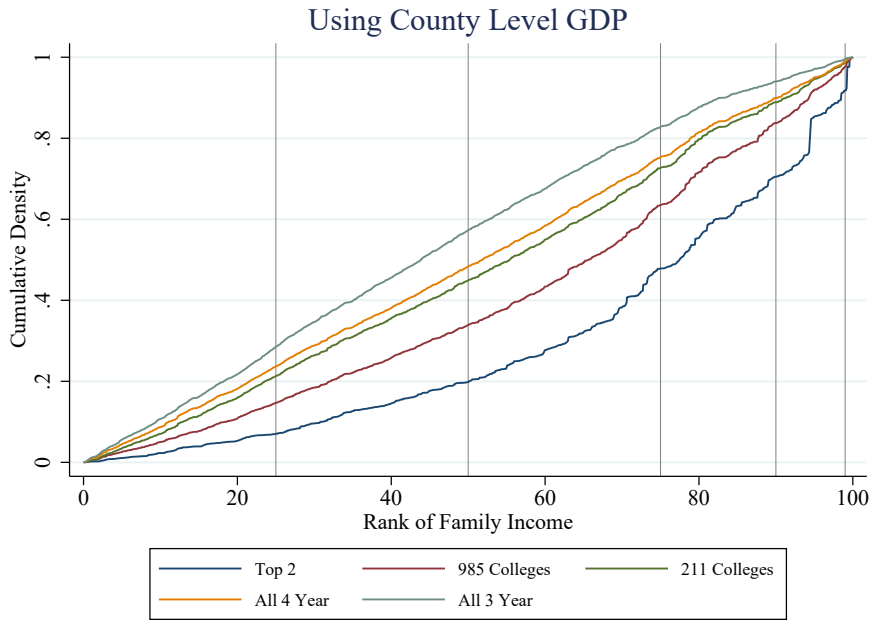
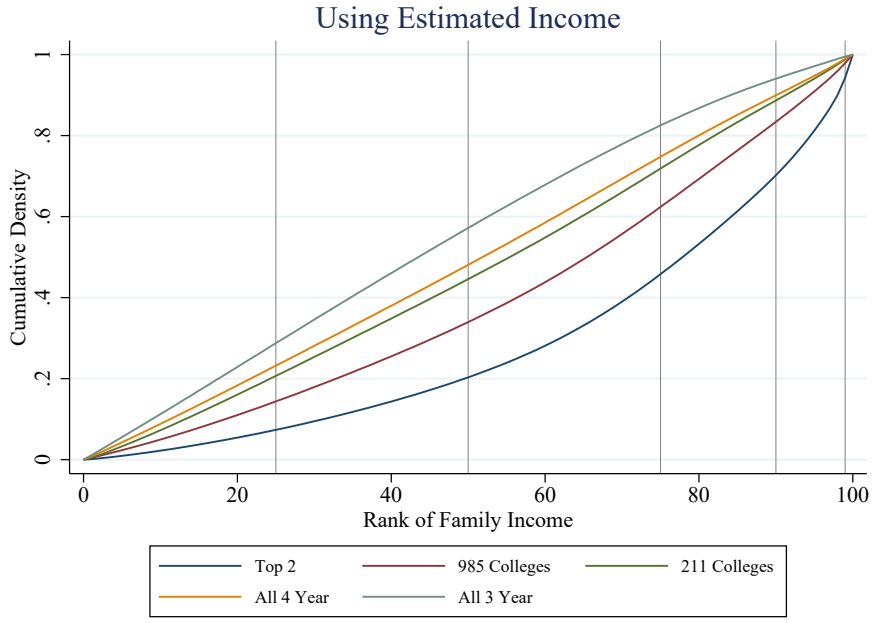


Figure 7: Cumulative Density of Student Income using County Level GDP and Estimation, 2010



(a) Using County Level GDP



(b) After Estimation

Table 1: Number of Observation in Admission Data and County Level Data Coverage

	4-Year Colleges		3-Year Colleges	
Year	Number of Students	Coverage	Number of Students	Coverage
2005	2185325	0.977	0	.
2006	2330805	0.978	0	.
2007	2479387	0.974	0	.
2008	2686034	0.972	3121472	0.967
2009	2640639	0.968	3000721	0.963
2010	2839312	0.968	2859637	0.961
2011	2787171	0.968	2525222	0.968

A Data notes

Our college admission data covers all students admitted to a four-year college through NCEE in China from 2005 to 2007 and all students admitted to a three-year or four-year college through NCEE in China from 2008 to 2011, except a few provinces. The following groups are missing in our dataset:

- students from Jiangsu Province admitted to any colleges between 2009 and 2011;
- students from Zhejiang Province admitted to four-year colleges in 2011;
- students from Zhejiang, Jiangxi, and Hainan Provinces admitted to three-year colleges in 2011.

The coverage of county level GDP and population panel is detailed in Table 2 and 3.

Table 2: Coverage of County Level Datasets, 4 Year College Admits

Year	Student Number	GDP and Pop	GDP	Pop	Raw[1]
2005	2185325	0.977	.979	0.979	0.866
2006	2330805	0.978	.98	0.980	0.879
2007	2479387	0.974	.976	0.976	0.870
2008	2686034	0.972	.975	0.981	0.870
2009	2640639	0.968	.971	0.971	0.869
2010	2839312	0.968	.971	0.971	0.888
2011	2787171	0.968	.972	0.976	0.772
Total	2564096	0.972	.975	0.976	0.859

(a) Overall Coverage of County Level Panel

Year	CEIC[2]	EPS	Pro Stat Yrbk	WIND	Stat Com
2005	.626	0.700	.685	0.025	0.029
2006	.623	0.709	.689	0.024	0.028
2007	.607	0.692	.674	0.023	0.035
2008	.613	0.671	.666	0.023	0.031
2009	.606	0.670	.668	0.041	0.035
2010	.62	0.637	.617	0.047	0.021
2011	.559	0.475	.446	0.043	0.005
Total	.608	0.651	.635	0.032	0.026

(b) Coverage of County Level GDP by Source

Year	CEIC	EPS	Pro Stat Yrbk	WIND	Stat Com
2005	.646	0.785	.654	0.010	0.054
2006	.646	0.784	.654	0.011	0.052
2007	.631	0.778	.654	0.012	0.057
2008	.641	0.787	.63	0.016	0.047
2009	.623	0.793	.626	0.017	0.031
2010	.625	0.800	.619	0.016	0.024
2011	.619	0.806	.589	0.017	0.004
Total	.633	0.790	.633	0.014	0.038

(c) Coverage of County Level Population by Source

[1] See table 3 notes.
[2] See table 3 notes.

Table 3: Coverage of County Level Datasets, 3 Year College Admits

Year	Student Number	GDP and Pop	GDP	Pop	Raw[1]
2008	3121472	0.967	.97	0.976	0.881
2009	3000721	0.963	.966	0.966	0.884
2010	2859637	0.961	.965	0.965	0.897
2011	2525222	0.968	.968	0.968	0.826
Total	1643865	0.965	.967	0.969	0.872

(a) Overall Coverage of County Level Panel

Year	CEIC[2]	EPS	Pro Stat Yrbk	WIND	Stat Com
2008	.667	0.735	.74	0.019	0.027
2009	.66	0.741	.749	0.040	0.028
2010	.674	0.711	.704	0.045	0.013
2011	.625	0.543	.527	0.038	0.003
Total	.657	0.682	.68	0.036	0.018

(b) Coverage of County Level GDP by Source

Year	CEIC	EPS	Pro Stat Yrbk	WIND	Stat Com
2008	.689	0.821	.654	0.011	0.036
2009	.674	0.832	.651	0.012	0.023
2010	.678	0.835	.623	0.012	0.016
2011	.672	0.844	.623	0.012	0.002
Total	.678	0.833	.638	0.012	0.019

(c) Coverage of County Level Population by Source

[1] Raw refers to using only directly available county data, excluding data for remaining counties in same city calculated by subtracting available county data from city level data.

[2] CEIC combines data from Municipal and National Bureau of Statistics, EPS China Regional Economy database collects data from China Statistical Yearbook for Regional Economy and China County Statistical Yearbook, Pro Stat Yrbk stands for provincial statistical yearbooks, WIND stands for WIND database, and Stat Com contains data collected from county level Statistical Communiqué collected from http://www.tjcn.org/tjgb/.