Family Income and Educational Returns in China, 2010-2020: Higher Returns Favor the Wealthy, but Gaps Hold Steady*

Yisu Hou

December 4, 2024

Economic returns to education are a key driver of socioeconomic mobility, yet they are not distributed equally across all income groups. Using data from the China Family Panel Studies (CFPS) between 2010 and 2020, this paper analyzes how family income influences the financial benefits individuals gain from education. The analysis reveals that individuals from wealthier families consistently enjoy significantly higher income returns from education compared to their lower-income counterparts, though this gap did not widen over the decade studied. These findings highlight persistent inequality in education's economic impact, emphasizing the need for policies that ensure education serves as a more effective equalizer in society.

1 Introduction

Since the economic reform in China in 1978, China's GDP per capita has increased more than 29 times (NBSC 2020). However, this period of rapid economic growth has been accompanied by greater socioeconomic inequality. The increased stratification, along with a number of China specific policies, has pushed the GINI coefficient close to 0.5, and various studies have demonstrated that the majority of China's wealth is concentrated in the hands of a few (Yu Xie and Jin 2015).

In examining the factors contributing to China's increasing inequality, a multitude of social scientists highlight China's education system. The correlation between socioeconomic status and education has been widely scrutinized. According to Esping-Andersen, "in modern capitalist society, the two main exclusionary devices through which the bourgeoisie structures

---

*Code and data are available at: https://github.com/YisuHou1/Educational_Returns_Analysis.

1

and perpetuates itself as a class, are first, those surrounding the institutions of property, and second, academic or professional qualifications and credentials." (Esping-Andersen 1982)

In China, educational attainment directly corresponds to future earnings. Research by Byron and Manaloto showed an approximate 4 percent increase in earnings for each year of schooling (Byron and Manaloto 1990), and Wu and Xie identified a rate of 4.7 percent (Wu and Xie 2003). Jansen and Wu documented a significant rising trend of returns on education during China's economic reform period from 1978 to 2005 (Jansen and Wu 2012). Concurrently, China significantly expanded its higher education sector, raising the gross enrollment rate from 9.8 percent in 1998 to 24.2 percent in 2009 (X. Wang and Liu 2011). As of 2022, the enrollment rate stands at 59.7 percent, indicating that China has fully transitioned from elite to mass higher education. Although increased access to higher education and higher returns on education are generally believed to foster social mobility, China's inequality has worsened. To elucidate this paradoxical outcome, researchers found that the new educational opportunities are predominantly accessed by privileged groups, such as individuals with higher social status, males, and urban populations (Ou and Hou 2019).

Previous research has also investigated the reasons behind the educational gap between high and low-income individuals. In the 20th century, many theorists proposed theories examining the link between socioeconomic status and educational achievement. Some suggest that income-based segregation places low-income children in overcrowded classrooms with less adequately trained educators. Others argue that teachers may harbor negative attitudes towards low-income children, anticipating they will perform poorly academically, thereby creating a self-fulfilling prophecy (Bond 1981). More recently, Duncan and Murnane suggested that the disparity in child enrichment expenditure influences academic performance, giving children from high-income families an advantage (Duncan and Murnane 2014).

However, while the relationship between financial status and educational attainment is relatively well established, there is a dearth of research, especially in China, to explore the link between family financial status and the rate of return from education. This is an essential connection, as it signifies how effectively education translates into its ultimate outcome: income. The aim of this research is to fill this knowledge gap by examining whether individuals from different socioeconomic backgrounds in the Chinese population receive the same level of returns from education.

The estimand of this investigation is the disparities in educational returns among individuals from different family wealth quartiles during the decade 2020 to 2020, measured in the percent income increase per year of education received.

To arrive at a conclusion, national family and individual data from China Family Panel Studies (ISSS 2010) from the year 2010 to 2020 were extracted. For the data in each CFPS iteration, Propensity Score Matching (PSM) was utilized to balance the control variables across different family wealth quartiles, and separate polynomial regressions were fit to identify the disparities of educational returns across quartiles. It was found that throughout the decade, individuals from wealthier families possessed a significant advantage over individuals from less wealthy

families, with disparities reaching more than 3% income increase per year of education between some quartiles.

The body of the paper contains the data, model, results, and discussion sections. In order, Section 2 contains an overview of the data, the variables used, and the programming tools used in my analysis and interpretation of the data. Section 3 describes the model in detail and provides justification for it. Section 4 contains a run-down and explanation of the results and findings from the model, and Section 5 contains a discussion of the implications and limitations of the result, among others. The appendices contain a detailed examination of the methodologies of CFPS, as well as aspects of the data and analysis excluded from the body.

2 Data

2.1 Overview

I used the statistical programming language R (R Core Team 2023) for the graphing, analysis and presentation of the project as a whole. Caret (Kuhn and Max 2008), glmnet (Friedman, Tibshirani, and Hastie 2010; Simon et al. 2011; Tay, Narasimhan, and Hastie 2023), nnet (Venables and Ripley 2002), tidyverse (Wickham et al. 2019), car (Fox and Weisberg 2019) MatchIt (Ho et al. 2011) and arrow (Richardson et al. 2024) were used directly in the analysis of the data. Stringr (Wickham 2023), dplyr (Wickham et al. 2023), gt (Iannone et al. 2024), knitr (Yihui Xie 2023), kableExtra (Zhu 2024) and styler (Müller and Walthert 2023) were used in the presentation and/or styling of the data, graphs, and paper as a whole.

All data utilized in this study were sourced from various iterations of the China Family Community Survey (ISSS 2010). The CFPS is a nationwide community survey funded by 985 Program of Peking University and carried out by the Institute of Social Science Survey of Peking University. CFPS encompasses approximately 15,000 families across 25 provinces and aims to tracks them permanently since 2010. Its main objective is to gather data across three dimensions: individuals, households, and communities. To achieve this, CFPS comprises four sub-surveys that thoroughly investigates families, adults, and children, gathering data for thousands of variables. Given its expansive and representative sample and a wide range of variables, CFPS has become a staple dataset for research in Chinese social sciences. Alternative data sources for this study, such as China General Social Survey (CGSS) and China Social Survey (CSS) are also reliable but do not provide the same width of variables, especially concerning family units. For the purposes of this research, data was extracted from CFPS for the years 2010, 2012, 2014, 2016, 2018, and 2020.

2.2 Measurement

2.3 Outcome variable

The variables presented in this section are variables in the polynomial regression model to identify returns to education, which is a coefficient of the model. This will be specified in Section 3.

The outcome variable of the model is the natural logarithm of individuals' hourly wages. This variable is not directly available in CFPS. It was calculated by dividing individuals' yearly work hours from their yearly income. Figure 1 shows its distribution in the aggregated survey data from all CFPS iterations between 2010 and 2020.
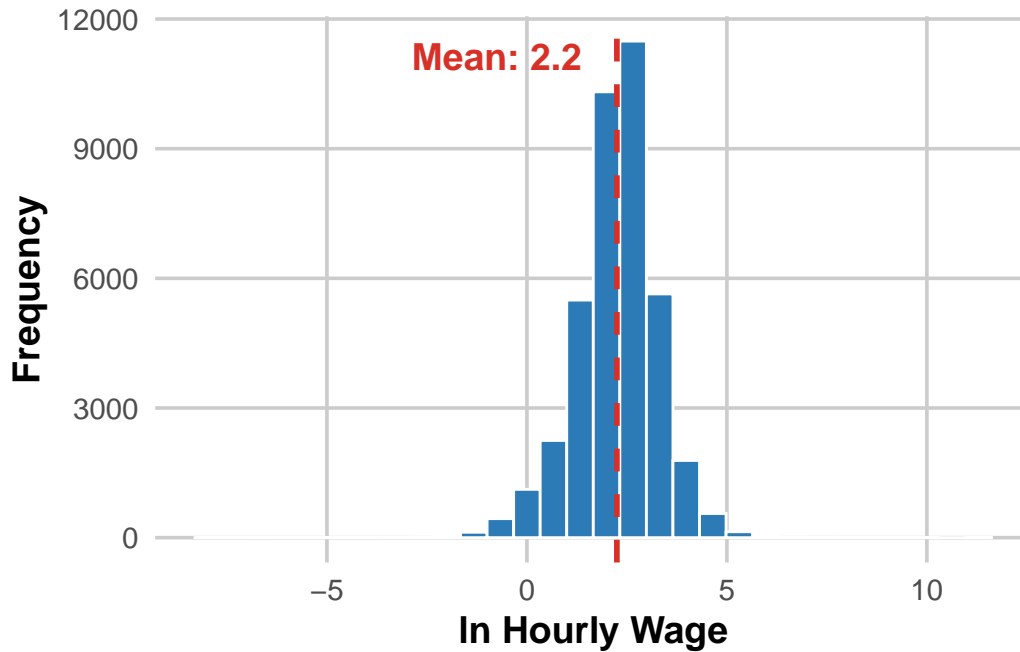

Figure 1: Natural Log of Individual Hourly Wage, CFPS2010 to CFPS2020

As displayed by Figure 1, the distribution of the natural log of hourly wage is close to a normal distribution, with a mean of 2.2 and most data points between 0 and 5. However, the range of values extends from around -9 to 12, representing individuals with extremely low and high income.

2.4 Predictor variables

2.4.1 Years of Education Received

Years of education is selected as the primary predictor in the model to identify economic returns to education. The theoretical framework for the model will be explained in Section 3.
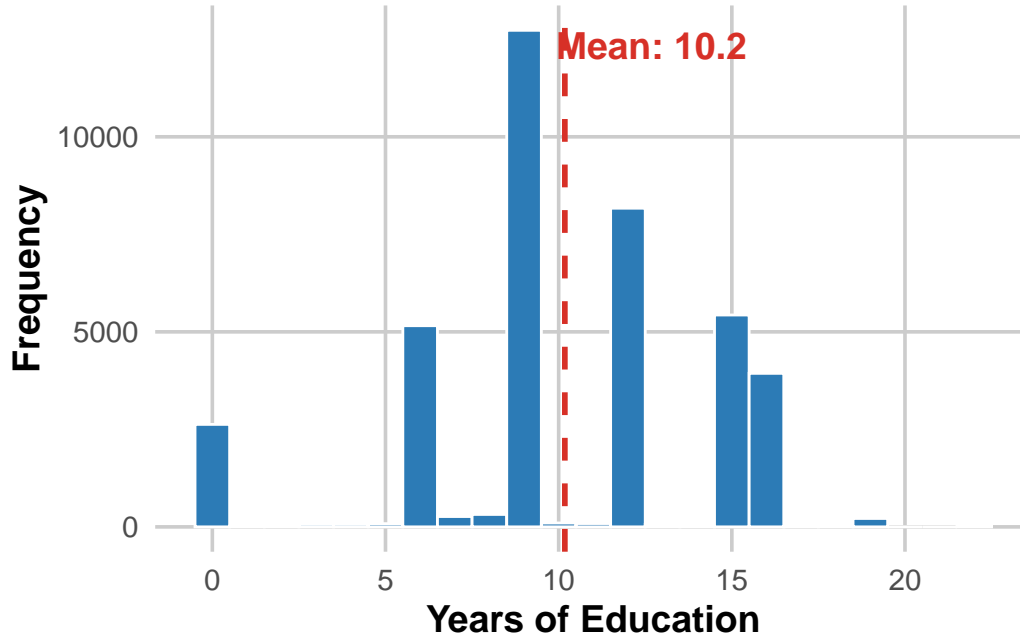

Figure 2: Years of Education Received, CFPS2010 to CFPS2020

As shown in Figure 2, most individuals in the aggregate dataset take very specific values, such as 9, 12, and 15. This is because most iterations of the CFPS record education received as the highest level of education completed. During the data cleaning process, they were translated to their corresponding number of years. For example, an individual who completed high school receives a 12 for years of education. I am aware that this could be highly inaccurate, as individuals may skip or repeat a grade. This is a limitation of the study that will be fully discussed in Section 5.

2.4.2 Gender

In light of the burgeoning prominence of the gender wage gap as a global sociological concern in recent decades, extensive research has been undertaken to discern the connection between gender and income. Scholars discovered that Chinese females earned 84.4 percent of male wages in 1988 and 82.5 percent in 1995 (Björn Gustafsson and Li 2000). Zhang identified a

38.23 percent wage gap using 2004 data, demonstrating an exacerbation of the issue over time (Zhang 2006).

Beyond merely influencing wage, studies have found a direct bearing of gender on educational returns, particularly in China. Gustafsson and Li found that in 1988, a college diploma raised a woman's income by 10.2 percent and a man's income by 8.9 percent compared to high school graduates (Björn Gustafsson and Li 2000). In 1995, the values changed to 20.8 percent for women 15.5 percent for men. A more recent study questioned the real magnitude of the gender-based difference in educational returns, as Chinese couples often pool income together, and many females have access to their spouse's income (Hannum, Zhang, and Wang 2013). Nevertheless, the role of gender as a determinant influencing wage and educational returns is well-established. Thus, it was included as a predictor.
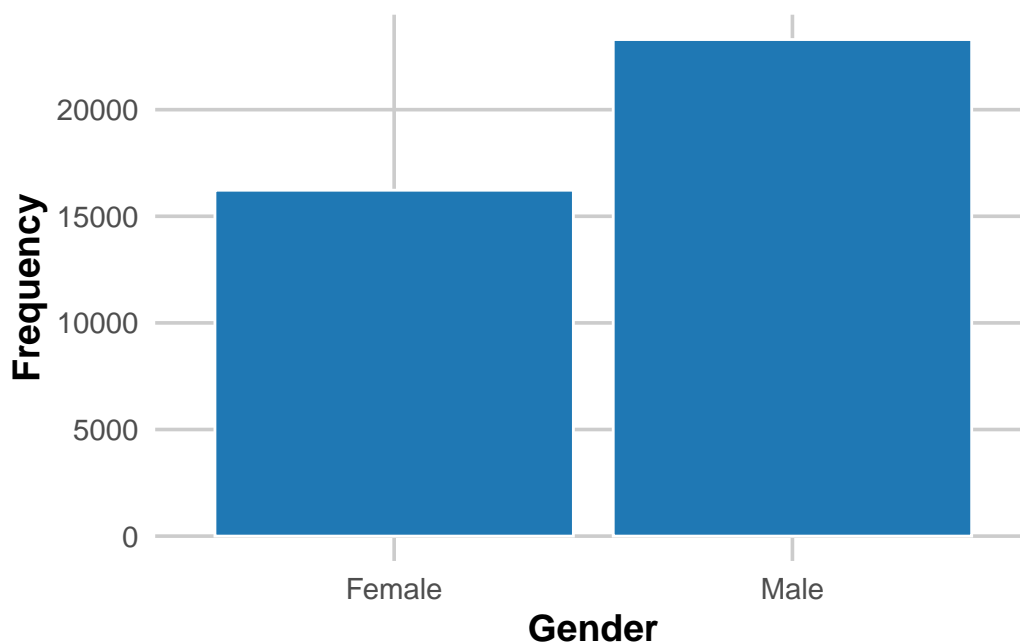

Figure 3: Gender of Respondent, CFPS2010 to CFPS2020

Figure 3 shows that there were significantly more male CFPS respondents than female respondents throughout 2010-2020.

2.4.3 Work Experience

Work experience is another key variable in the model for educational returns. CFPS does not directly provide data for work experience, so potential work experience was created as a substitute.

Potential work experience was deduced with an individual's age and their years of schooling. Established norms in labor economics prescribe the following age thresholds for labor market entry based on educational attainment: 17 for high-school dropouts, 19 for high-school graduates, 21 for high-school graduates with some college, and 23 for college graduates (Borjas 2003; Ottaviano and Peri 2012; Edo and Rapoport 2019). Thus, potential work experience was calculated by subtracting the corresponding value from an individual's age. This assumes that individuals directly enter the labor market after completing their education, which is a strong assumption.
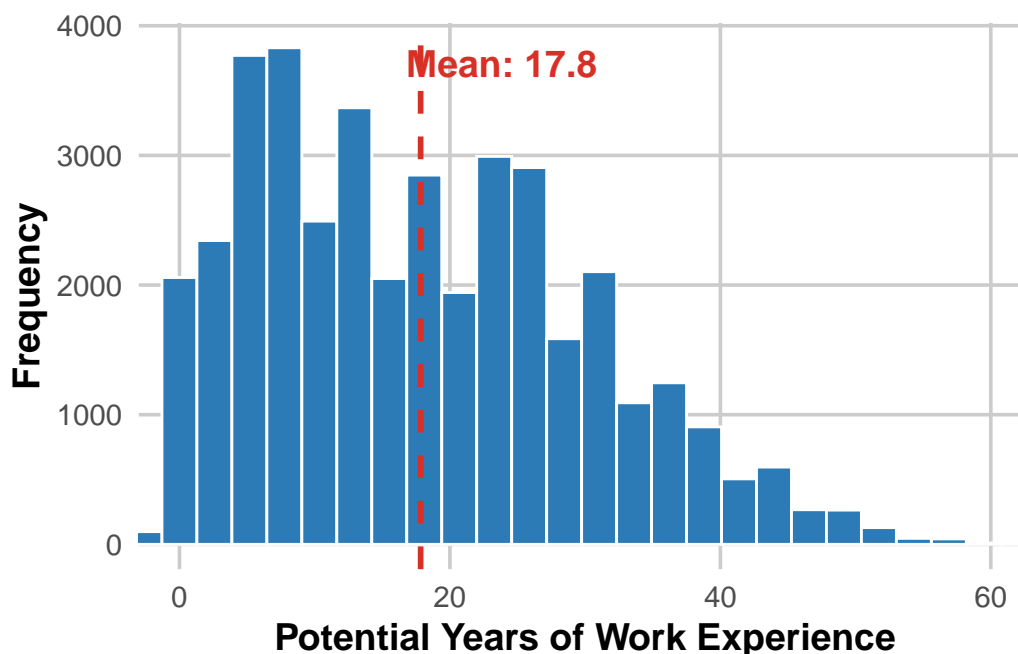

Figure 4: Potential Years of Work Experience, CFPS2010 to CFPS2020

Figure 4 shows that potential work experience is a right-skewed distribution with a mean of 17.8. This is expected, as it is unlikely for individuals to work for more than 40 years.

2.4.4 City Hukou Status

The Hukou system in China, a household registration mechanism, classifies individuals into either rural or urban demographics based on their birth registration locale. Individuals with urban Hukou and live in major cities enjoy the lion share of quality education and opportunities to pursue higher education. According to Tam and Jiang, during the recent period of higher education expansion, urban students' access to higher education rose sharply while rural students' access stagnated, showing the growing advantage of urban students due to unequal distribution of educational resources (Tam and Jiang 2015). The scholars explained

the widening gap by pointing to the expansion of vocational education, which were accessible to low-performing urban students who would not have had the opportunity to pursue higher education prior to the expansion. However, rural students were not provided with equal opportunities in vocational education.

How the Hukou system affects educational attainment is a complex topic that must be examined together with other phenomena in China. Local governments are responsible for educational institutions within its region, but rural governments only have a minimal proportion of the capital available to large cities; top universities are disproportionally located in major cities and are more available to individuals possessing a Hukou of those cities; migrant workers, or individuals who live in cities but do not hold a Hukou of the city they live in, do not have access to public schooling in the city and may only attend migrant schools with worse conditions and educational quality (C. Wang 2007).

Therefore, it is no surprise that urban Hukou holders have higher wages. Using data from 2005, scholars found a rural Hukou worker's wage per hour is only 64 percent of an urban Hukou worker's (Xing, Jia, and Li 2013).
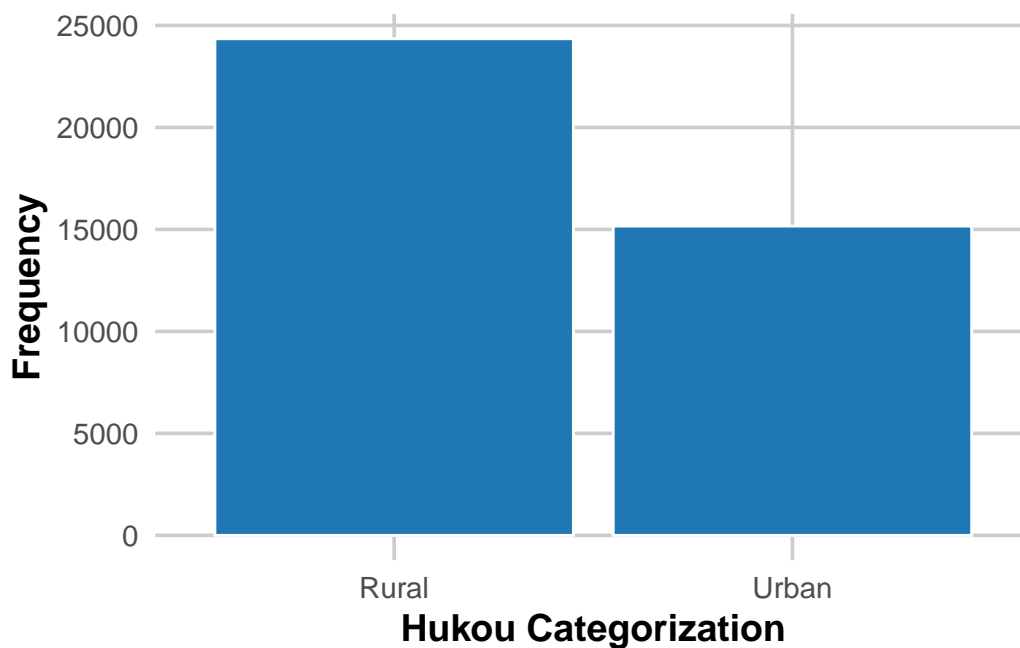

Figure 5: Hukou Status of Respondent, CFPS2010 to CFPS2020

Figure 5 shows that there are a greater number of rural Hukou holders than urban Hukou holders among CFPS respondents. This accurately reflects the overall demographics of China, where the rural population is larger than the urban population.

2.4.5 Marital Status

Studies have shown the magnifying effect of marriage on gendered income gaps, stemming from gender-based roles and expectations within familial settings. Wang and Wang demonstrated that unmarried Chinese men and women in 2015 had an average monthly income gap of 458.612 RMB, and married couples had a gap of 1398.80 RMB (Z. Wang and Wang 2020). Married couples with children under 18 had an income gap that is more than doubled compared to those without children. Mu and Xie explained the after-marriage changes by addressing that fathers acquire a confidence boost after the birth of their children and dedicate themselves in the pursuit for higher wages and a successful career (Mu and Xie 2016). On the other hand, mothers take the child-rearing role, accumulate less experience in the labor force, and choose less rigorous jobs to reserve energy for household work.
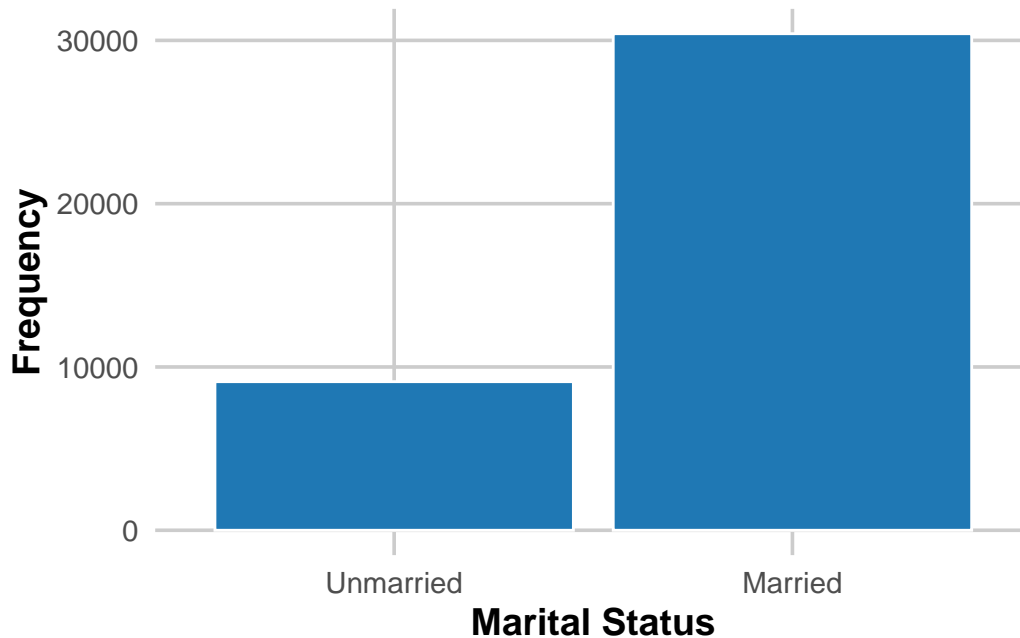

Figure 6: Marital Status of Respondent, CFPS2010 to CFPS2020

As displayed in Figure 6, most respondents of CFPS are married.

2.4.6 Communist Party Membership

Membership in the Communist Party of China (CCP) is frequently included in income related research within China (Björn Gustafsson and Li 2000; Hannum, Zhang, and Wang 2013; Mu and Xie 2016). Being a member of the CCP grants the opportunity for an individual to work in the government positions or state-owned enterprises that provide secure and well-paid jobs (Dickson 2014). Moreover, party status is also related to the level of education, as the

recruitment strategy of the CCP targets elite groups such as college and graduate students. This study followed established protocols and included CCP membership as a predictor, as its unique socio-political status in China necessitates its inclusion.
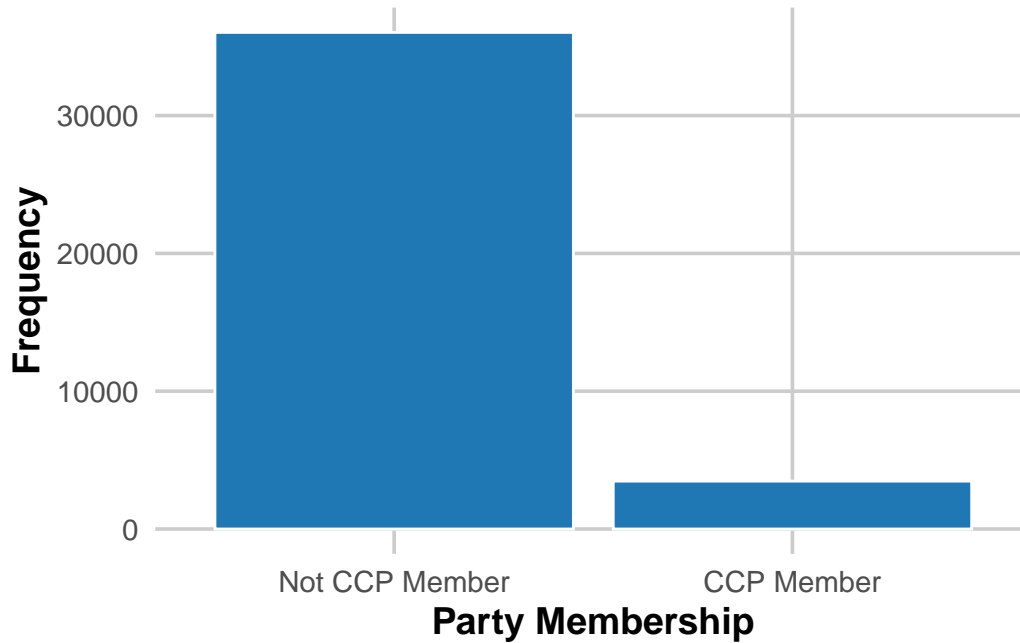

Figure 7: CCP Membership of Respondent, CFPS2010 to CFPS2020

Figure 7 reflects that CCP members are only make up a small proportion of the Chinese population.

2.4.7 Eastern Provinces

China's geographical East, encompassing 12 economically advanced provinces, is characterized by notable income disparities when compared to Central and Western regions. Scholars found that in 2005, Eastern populations not only had significantly greater income but also enjoyed doubled returns to education compared to other regions (Xing, Jia, and Li 2013). Though the values shifted towards the national mean in 2011, significant differences still existed. Recent research endeavors on income and educational returns have increasingly incorporated the geographical East as a control variable, acknowledging its substantial influence on the variables of interest (Dong, Yang, and Luo 2021).
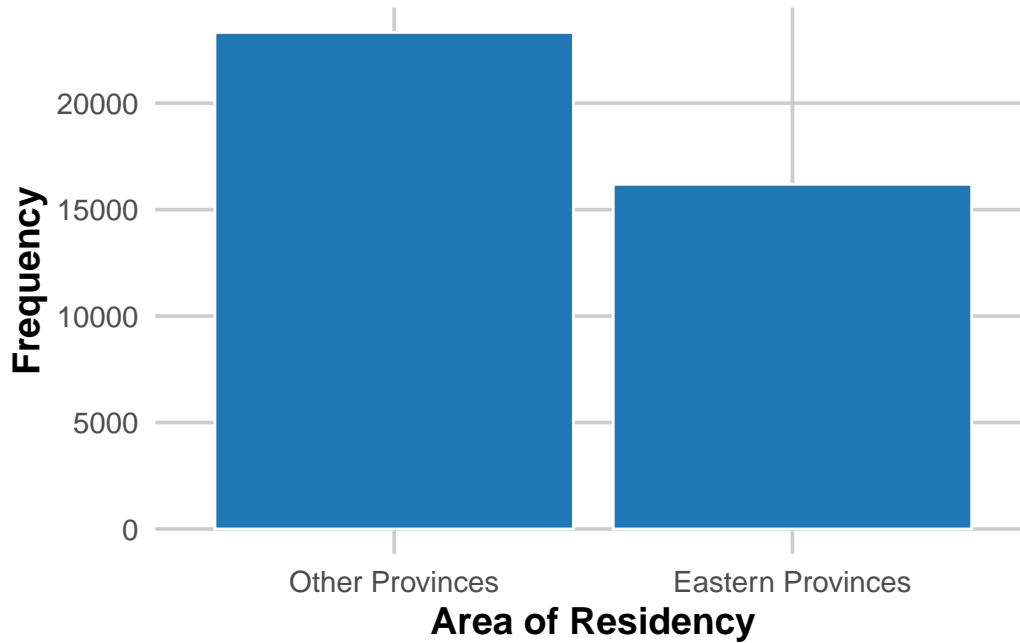
Figure 8: Location of Residence, CFPS2010 to CFPS2020

According to Figure 8, there exists a relatively balanced split between respondents residing in Eastern provinces and those living in other provinces. More individuals lived in non-Eastern provinces, counting to more than 20,000.

2.4.8 Family Income Quartile

Family income quartile is not a variable directly incorporated in the regression model. Rather, data was split into subsets according to family income quartiles and regressed to find the educational returns of each quartile in comparison with one another. This is the focal variable of the study. Within the CFPS's family economics sub-survey, there exists a variable listing family income quartiles. For this study, family economics datasets were merged with the adult individuals dataset of the same CFPS iteration by matching personal IDs, thus deriving a corresponding family income quartile value for each adult participant.
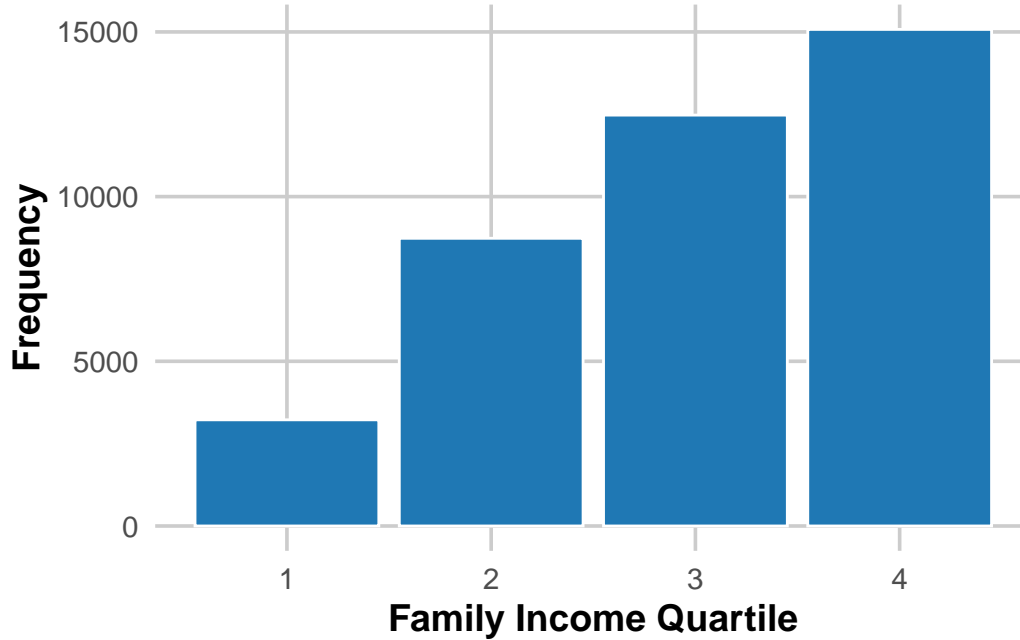
Figure 9: Family Income Quartiles, CFPS2010 to CFPS2020

The distribution shown in Figure 9 is counter intuitive, as there are more individuals who belong in upper quartiles. This does not suggest that CFPS uses an unrepresentative sample. The data of individuals in upper family income quartiles are relatively more complete, thus more data points are kept after the data cleaning process. Despite the relative lack of data for the lower quartiles, the number of data points is still sufficient for this study.

3 Model

The goal of my modelling strategy is twofold. First, I must split the data from each CFPS iteration into family income quartiles and adjust them to be directly comparable with one another. Second, I wish to calculate each quartile's returns to education, in terms of percent income increase per year of education, to complete the comparison.

To accomplish the first task, Propensity Score Matching was utilized.

Historically, research on returns to education primarily relied on directly calculating the rate of return using the Mincer Earnings Function. They incorporated control variables into the regression to account for their influence on income (Byron and Manaloto 1990; Zhang 2006). When desiring to make group comparisons, scholars generally subdivided data based on specific variables of interest such as gender, location, and industry. They then performed direct regressions to contrast the coefficients Mu and Xie (2016).

While multivariate regression effectively captures the impact of control variables, its functionality hinges on the linear relationships between the outcome and covariates. Dehejia and Wahba revealed that regressing groups with significant variances in the mean values of covariates could yield misleading results for the variable of interest, potentially overstating the differences between groups (Dehejia and Wahba 2002). Propensity Score Matching (PSM) addresses this issue by isolating the treatment effect. Therefore, this study employed a combined method: first balancing covariates using PSM, then executing regressions to adjust for residual differences.

For this study, data from each CFPS iteration were divided into four subsets based on family income quartiles. PSM was then applied to match each subset with others, resulting in six pairs of subsets that exhibited different treatment variable values (family income quartiles) but had similar means for control variables. The matching was conducted using the nearest-neighbor approach with $caliper = 0.1sd$, which is more strict compared to the generally recommended 0.2 threshold for caliper width (Austin 2011). This choice was intentional, as reducing the caliper width improved the covariate balance without significantly reducing the number of data points. Given that lower family income quartiles had fewer data points, match ratios were adjusted based on the counts within the quartiles being matched.

To show the effects of PSM, Table 1 provides a summary of balance for the match between the Quartile 1 and the Quartile 2 subset of CFPS 2020. The match was executed using caliper = 0.1 and a ratio of 2, matching 1041 Quartile 2 instances out of 1324 total instances with all 541 Quartile 1 instances. This matching procedure was repeated 6 times to compare each quartile with every other quartile for all 6 CFPS iterations from 2010 to 2020.

Table 1: Summary of Balance for Quartile 1 vs Quartile 2 Match, 2020 Data

Variables	Means.Q1	Means.Q2	Std.Mean.Diff	Means.Q1.Matched	Means.Q2.Matched	Std.Mean.Diff.M
distance	0.3016	0.2854	0.3099	0.3016	0.3009	0.0139
Gender	0.5712	0.5650	0.0125	0.5712	0.5675	0.0075
Work exp	14.8817	14.1767	0.0866	14.8817	14.7902	0.0112
City Hukou	0.1072	0.1971	-0.2907	0.1072	0.1100	-0.0090
Marital Status	0.7375	0.7515	-0.0318	0.7375	0.7366	0.0021
CCP	0.0240	0.0219	0.0139	0.0240	0.0222	0.0121
East	0.2810	0.3112	-0.0672	0.2810	0.2810	0.0000

For the second task, a multivariate regression model was created using the theoretical framework of the Mincer Earnings Function (Mincer 1974). The function depicts wage as a function of schooling and work experience. Its original form is:

$$ln(w) = f(s, x) = ln(w_0) + \rho s + \beta_1 x + \beta_2 x^2$$

Where w is hourly wage, w_0 is the wage level of an individual with 0 years of schooling and work experience, s is years of schooling, x is work experience, and ρ, β_1, and β_2 are coefficients that can be viewed as the rates of return to schooling and work experience.

Numerous researchers have since tailored the Mincer equation by incorporating potential confounders, ensuring the mitigation of biased outcomes. A seminal study analyzing Chinese educational returns used the function:

$$ln(Y) = \alpha_0 + \alpha_1 Edu + \alpha_2 Exper + \alpha_3 Exper^2 + \sum \lambda_j X + \mu$$

Which adds X that represents control variables such as age, gender, and geographical location (Zhang 2006). Analogous modifications can be observed globally, although with variable inclusions that reflect regional societal and cultural differences. For example, an Indian study choose to include caste, relationship to the head of household, and the ability to speak and write English in the equation as control variables (Bhandari and Bordoloi 2006). This research also leaned on a modified version of the Mincer Earnings Function for its calculations on educational returns.

It is acknowledged that the regression models may be inconsistent due to the inconsistencies in the variables and measurements of different CFPS iterations. The incorporation of calculated variables is an additional factor impacting the models' ability to calculate returns to education.

3.1 Model set-up

3.1.1 Multivariate regression models

Employing PSM-matched subsets, the educational returns for different family income quartile groups were computed using the modified Mincer Function:

$$ln(w) = \alpha_0 + \alpha_1 Edu + \alpha_2 Exp + \alpha_3 Exp^2 + \alpha_4 Gender + \alpha_5 Hukou + \alpha_6 Marriage + \alpha_7 CCP + \alpha_8 East + \mu$$

With w representing hourly wage, α_1 representing returns to education and other coefficients representing the effects of previously identified covariates on the natural log of income.

Next, PSM-matched subset pairs were separated by family income quartiles and regressed independently to identify the differences between educational returns and the coefficients of covariates.

Finally, the Delta coefficient α_1 from all paired comparisons across every CFPS iteration were compiled together to identify cross-quartile education return disparities over time.

As a test of model robustness, another set of regressions were ran without PSM to verify the overall significance of educational returns and the consistent pattern observed between quartiles. This will be discussed further in Section 4.

4 Results

4.1 Cross-quartile disparities in Educational Returns

Table 2: Cross-Quartile Percentage Disparities of Returns to Education Over Time

CFPS.Year	Q1vsQ2	Q1vsQ3	Q1vsQ4	Q2vsQ3	Q2vsQ4	Q3vsQ4	Cross.Sectional.Mean
2020	0.9482	1.9092	4.1670	1.2350	3.1601	2.2804	2.2833
2018	0.2441	2.3463	3.9833	1.7767	3.3524	3.7010	2.5673
2016	0.7665	1.2655	4.5967	1.2912	3.9390	3.1417	2.5001
2014	-2.3488	-1.1069	0.6399	0.2706	3.3242	3.1945	0.6623
2012	1.2055	0.9527	3.1572	-0.0212	3.7233	4.4302	2.2413
2010	0.4988	3.0412	5.6921	2.1727	4.9265	2.1920	3.0872
Temporal mean	0.2191	1.4013	3.7060	1.1208	3.7376	3.1566	

Note: In 2012, educational returns for Q1, Q2, and Q3 were insignificant. In 2010, educational returns for Q1 and Q2 were insignificant.

As Table 2 shows, in most CFPS iterations, individuals from higher family income quartiles demonstrated a significant advantage in terms of returns to education compared to their counterparts in other quartiles. The greatest differences were observed when comparing the highest income quartile (Q4) to other quartiles. Over all CFPS iterations, individuals in the top quartile benefitted from 3.71% more gains in income from an additional year of education compared to the lowest quartile, 3.74% more than the second quartile, and 3.16% more than the third quartile on average. Comparatively, the differences between the other quartiles were less pronounced.

Upon evaluating cross-sectional means, no discernible upward or downward trend was identified. Other than the outlier year 2014, individuals from higher quartiles experienced around 2 to 3 percent greater educational returns on average compared to individuals from relatively lower quartiles. Thus, a general inference can be made: while those in the higher family income quartiles consistently enjoy significant benefits in terms of returns to education, this advantage has not expanded over time.

4.2 Robustness testing

Table 3: Quartiles' Percentage Returns to Education Over Time Without PSM

CFPS.Year	Q1.Returns	Q2.Returns	Q3.Returns	Q4.Returns
2020	4.9507***	5.8130***	7.3284***	10.3891***
2018	4.1730**	4.0239***	6.1466***	9.1724***
2016	2.5671*	3.0501***	4.1665***	6.9051***
2014	4.4464**	2.5357**	2.8027***	6.3410***
2012	-1.0031	0.0590	0.4928	4.4870***
2010	0.1803	1.1116	3.9577***	8.0710***

Significant codes: * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$

When directly regressing without matching, the significance of educational returns coefficients remained the same; the coefficients for Q1, Q2 and Q3 of CFPS2012 and the coefficients for Q1 and Q2 of CFPS2010 are the only ones insignificant at the 5% level. It is uncertain why educational returns are insignificant for these family income quartiles, both in the matched data and the unmatched data. It might be caused by the legacy data collection methods or the changes in variable recording that occurred after the 2012 iteration. Despite that, the outcome of the robustness test demonstrates the model's consistency.

5 Discussion

5.1 Takeaways

This paper investigates how family income influences the economic returns to education in China from 2010 to 2020 using data from the China Family Panel Studies (CFPS). Propensity Score Matching (PSM) and multivariate regression models were utilized to analyze cross-quartile disparities.

The study reveals that individuals from wealthier families consistently experience significantly higher income returns from education compared to those from lower-income families. These disparities are particularly pronounced when comparing the highest income quartile to the other quartiles. However, the gap in returns has remained stable over the studied decade, suggesting persistent inequality without evidence of widening.

From this research, we learn about the role of socioeconomic background in shaping the economic benefits of education in China. The finding that the highest income quartile produces significantly greater disparities than other quartiles hints that being a member of the middle class does not necessarily imply better social mobility compared to the lower class, but being in the upper class produces greater economic benefits for the working individual. On the large scale, it highlights a fundamental limitation in the equalizing potential of education, as the higher financial returns to education predominantly favor individuals from wealthier families.

This underscores the need for targeted policies to address these disparities and enhance education's role as a tool for socioeconomic mobility; merely expanding higher education may not be sufficient.

5.2 Weaknesses and next steps

This research has a number of limitations. First, the data from CFPS has poor completion rates for some of the variables. Many individuals from lower income quartiles did not respond to questions such as weekly hours worked or income. This caused the data used in this research to include more individuals from higher quartiles, increasing bias. The second limitation is that the sample changed over time as individuals responded differently in different CFPS iterations. One of the greatest benefits of CFPS is that it follows the same pool of families over time, but the data for this study did not include the same set of individuals throughout different CFPS iterations because respondents left different questions unanswered in different years. Consequently, the sample sizes and the individuals making up the samples were different for each year, making the temporal analysis less conclusive. Finally, this research calculated returns to education using the Miner equation, which has a set of limitations. Besides the problem of linearity, it also assumes that all covariates influencing income are included in the model, which might not be the case. Additionally, the model only provides a general statement on the overall returns to education of different family income quartiles and does not provide insight to the reasons behind the differences.

For further research, one may execute further divisions in terms of educational stages and examine whether the difference of returns for various educational stages can explain the difference in overall educational returns between income groups.

Appendix

.1 Survey methodology overview and evaluation

.2 Idealised methodology

References

Austin, P. C. 2011. "Optimal Caliper Widths for Propensity-Score Matching When Estimating Differences in Means and Differences in Proportions in Observational Studies." *Pharmaceutical Statistics* 10 (2): 150–61. https://doi.org/10.1002/pst.433.

Bhandari, L., and M. Bordoloi. 2006. "Income Differentials and Returns to Education." *Economic and Political Weekly* 41 (36): 3893–3900. https://www.jstor.org/stable/4418680.

Bond, G. C. 1981. "Social Economic Status and Educational Achievement: A Review Article." *Anthropology & Education Quarterly* 12 (4): 227–57. https://www.jstor.org/stable/3216573.

Borjas, George J. 2003. "The Labor Demand Curve Is Downward Sloping: Reexamining the Impact of Immigration on the Labor Market." *The Quarterly Journal of Economics* 118 (4): 1335–74. https://doi.org/10.1162/003355303772812313.

Byron, R. P., and E. Q. Manaloto. 1990. "Returns to Education in China." *Economic Development and Cultural Change* 38 (4): 783–96. https://www.jstor.org/stable/1153971.

Dehejia, R. H., and S. Wahba. 2002. "Propensity Score-Matching Methods for Nonexperimental Causal Studies." *The Review of Economics and Statistics* 84 (1): 151–61. https://www.jstor.org/stable/3211745.

Dickson, Bruce J. 2014. "Who Wants to Be a Communist? Career Incentives and Mobilized Loyalty in China." *The China Quarterly*, no. 217: 42–68.

Dong, Yanfang, Zebing Yang, and Changfu Luo. 2021. "   ."   10: 58–70.

Duncan, G. J., and R. J. Murnane. 2014. "Growing Income Inequality Threatens American Education." *The Phi Delta Kappan* 95 (6): 8–14. https://www.jstor.org/stable/24374505.

Edo, Anthony, and Hillel Rapoport. 2019. "Minimum Wages and the Labor Market Effects of Immigration." *Labour Economics* 61: 101753. https://doi.org/10.1016/j.labeco.2019.101753.

Esping-Andersen, G. 1982. "Review of Marxism and Class Theory: A Bourgeois Critique, by f. Parkin." *Acta Sociologica* 25 (1): 96–98. http://www.jstor.org/stable/4194388.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Friedman, Jerome, Robert Tibshirani, and Trevor Hastie. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1–22. https://doi.org/10.18637/jss.v033.i01.

Gustafsson, Björn, and Shi Li. 2000. "Economic Transformation and the Gender Earnings Gap in Urban China." *Journal of Population Economics* 13 (2): 305–29.

Gustafsson, B., and W. Zhong. 2000. "How and Why Has Poverty in China Changed? A Study Based on Microdata for 1988 and 1995." *The China Quarterly* 164: 983–1006. http://www.jstor.org/stable/655923.

Hannum, E., Y. Zhang, and M. Wang. 2013. "Why Are Returns to Education Higher for Women Than for Men in Urban China?" *The China Quarterly* 215: 616–40. https://www.jstor.org/stable/23510805.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2011. "MatchIt: Nonpara-

metric Preprocessing for Parametric Causal Inference." *Journal of Statistical Software* 42 (8): 1–28. https://doi.org/10.18637/jss.v042.i08.

Iannone, Richard, Joe Cheng, Barret Schloerke, Ellis Hughes, Alexandra Lauer, and JooYoung Seo. 2024. *Gt: Easily Create Presentation-Ready Display Tables*. https://gt.rstudio.com.

ISSS. 2010. "China Family Panel Studies." Peking University. https://www.isss.pku.edu.cn/cfps/en/index.htm.

Jansen, W., and X. Wu. 2012. "Income Inequality in Urban China, 1978–2005." *Chinese Sociological Review* 45 (1): 3–27. https://doi.org/10.2753/CSA2162-0555450101.

Kuhn, and Max. 2008. "Building Predictive Models in r Using the Caret Package." *Journal of Statistical Software* 28 (5): 1–26. https://doi.org/10.18637/jss.v028.i05.

Mincer, J. A. 1974. "The Human Capital Earnings Function." In *Schooling, Experience, and Earnings*, 83–96. National Bureau of Economic Research, Inc.

Mu, Z., and Y. Xie. 2016. "'Motherhood Penalty' and 'Fatherhood Premium'? Fertility Effects on Parents in China." *Demographic Research* 35: 1373–1410. https://www.jstor.org/stable/26332115.

Müller, Kirill, and Lorenz Walthert. 2023. *Styler: Non-Invasive Pretty Printing of r Code*. https://github.com/r-lib/styler.

NBSC. 2020. *2020   *. China Statistics Press.

Ottaviano, Gianmarco I. P., and Giovanni Peri. 2012. "Rethinking the Effect of Immigration on Wages." *Journal of the European Economic Association* 10 (1): 152–97. https://doi.org/10.1111/j.1542-4774.2011.01052.x.

Ou, D., and Y. Hou. 2019. "Bigger Pie, Bigger Slice? The Impact of Higher Education Expansion on Educational Opportunity in China." *Research in Higher Education* 60 (3): 358–91. https://www.jstor.org/stable/45180384.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://github.com/apache/arrow/.

Simon, Noah, Jerome Friedman, Robert Tibshirani, and Trevor Hastie. 2011. "Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent." *Journal of Statistical Software* 39 (5): 1–13. https://doi.org/10.18637/jss.v039.i05.

Tam, T., and J. Jiang. 2015. "Divergent Urban-Rural Trends in College Attendance: State Policy Bias and Structural Exclusion in China." *Sociology of Education* 88 (2): 160–80. https://www.jstor.org/stable/43743433.

Tay, J. Kenneth, Balasubramanian Narasimhan, and Trevor Hastie. 2023. "Elastic Net Regularization Paths for All Generalized Linear Models." *Journal of Statistical Software* 106 (1): 1–31. https://doi.org/10.18637/jss.v106.i01.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with s*. Fourth. New York: Springer. https://www.stats.ox.ac.uk/pub/MASS4/.

Wang, C. 2007. "Education and Social Inequality in China: Elite Groups Perpetuating Their Privileged Status." *China Perspectives* 3 (71): 110–16. https://www.jstor.org/stable/

24053285.

Wang, X., and J. Liu. 2011. "China's Higher Education Expansion and the Task of Economic Revitalization." *Higher Education* 62 (2): 213–29. https://www.jstor.org/stable/41477865.

Wang, Zhaoping, and Wenli Wang. 2020. " ." 42: 99–112.

Wickham, Hadley. 2023. *Stringr: Simple, Consistent Wrappers for Common String Operations.* https://stringr.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation.* https://dplyr.tidyverse.org.

Wu, X., and Y. Xie. 2003. "Does the Market Pay Off? Earnings Returns to Education in Urban China." *American Sociological Review* 68 (3): 425–42. https://doi.org/10.2307/1519731.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r.* https://yihui.org/knitr/.

Xie, Yu, and Yongai Jin. 2015. "Household Wealth in China." *Chinese Sociological Review* 47 (3): 203–29. https://doi.org/10.1080/21620555.2015.1032158.

Xing, Chunbing, Shuyan Jia, and Shi Li. 2013. " ." 11: 114–26.

Zhang, Chewei. 2006. " : ' ' ." 12: 59–70.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax.* http://haozhu233.github.io/kableExtra/.