

Online Appendix for: “How Does Caste Affect Entrepreneurship? Birth versus Worth”

Sampreet Singh Goraya*

Abstract

This appendix contains supplemental material for the article “How Does Caste Affect Entrepreneurship? Birth versus Worth”

*Department of Economics, Stockholm School of Economics. Email: sampreet.goraya@hhs.se.

Contents

A Data Appendix	2
A.1 MSME Dataset	2
A.2 Economic Census 2005	3
A.3 ASI-NSS 2006	3
A.4 Household Surveys 2004-05	4
A.5 All India Debt and Investment Survey	4
B Stylized Facts	5
B.1 Fact 1	5
B.2 Fact 2	7
B.3 Fact 3	8
C Model	17
C.1 Income and Wealth Inequality	17

A Data Appendix

A.1 MSME Dataset

The MSME dataset is based on MSME sector which is defined by the Micro, Small and Medium Enterprise Development (MSMED) act of 2006, spans the non-agricultural enterprises of the economy that are below a certain threshold of size (size in terms of original value of investment in plant, machinery and equipment). The investment limit for enterprises engaged in the manufacturing or production of goods is Indian rupees (INR) 100 million whereas for those providing or rendering in services is INR 50 million.

In particular, the act notified the following enterprises, whether proprietorship, Hindu undivided family, association of persons, co-operative society, partnership or undertaking or any other legal entity, by whatever name called:- In case of enterprises engaged in manufacturing or production of goods pertaining to any industry specified in the First Schedule to the Industries (Development and Regulation) Act, 1951, as: (i) a micro enterprise, where the investment in plant and machinery does not exceed 2.5 million rupees, (ii) a small enterprise, where the investment in plant and machinery is more than 2.5 million but does not exceed 50 million rupees; or (iii) a medium enterprise, where the investment in plant and machinery is more than 50 million rupees but does not exceed 100 million rupees. In case of the enterprises engaged in providing or rendering of services, as: (i) a micro enterprise, where the investment in equipment does not exceed 1 million rupees; (ii) a small enterprise, where the investment in equipment is more than 1 million rupees but does not exceed 20 million rupees; or (iii) a medium enterprise, where the investment in equipment is more than 20 million rupees but does not exceed 50 million rupees.

According to the 4th MSME dataset of India 2006, the MSME sector accounts for 37% of the manufacturing output and 89% of the total employment in the manufacturing sector. The sector is estimated to employ about 59 million individuals in over 26.1 million units throughout the country. Further, 1.5 million (5.94%) are registered MSMEs and 24.5 million (94.06 %) are unregistered MSMEs that employ 16.62 % and 83.38 % of the workforce respectively.

Sample Survey of Unregistered MSME's: As explained in [MSME \(2009\)](#), sample survey was adopted in Unregistered Sector. To reach the enterprises of Unregistered Sector universe, Multi-Stage Stratified Sampling procedure was adopted. Villages in rural area and towns in urban area were the first stage units (FSU) of the survey programme. Selected enterprises in the selected villages/towns were the second stage units (SSU) of survey. The stratification of districts was done on the basis of total number of enterprises as per Economic census 2005. The FSUs were selected by sample random sampling without replacement, both for rural and urban areas. In the first stage, in the selected FSU (villages/towns), door to door visit was made to list out all the non-agricultural enterprises and MSME. The list so prepared was used for selection of enterprises for survey. In the second stage, MSMEs were selected out of the list of MSME prepared for each FSU. In villages, if the total number of MSME were less than 15, all of them were surveyed. In villages with more than 15 total numbers of MSME, selection was resorted to limit the survey of MSME to 15. In case of towns, if the total number of MSME were less than 20, all of them were surveyed. In towns with more than 20 total number of MSME, selection was resorted to limit the survey of enterprises to 20. Altogether 1.26 lakh

MSME were surveyed in Unregistered sector by the above process. The sample provide multipliers to compute summary statistics for the population of enterprises. To compute descriptive statistics in this paper, I always use these sampling multipliers.

In particular, the MSME dataset provides the geographical information, industry classification, balance sheet variables, and the caste of the owner and employees. The measure of capital stock in the data is the market value of fixed assets. The total wage bill includes salaries and wages, allowances, bonuses, and so on. The measure of output is gross value added. The amount of loan outstanding captures all the loans from formal and informal sources, where informal sources include local moneylenders, friends and relatives.

Winsorization: The financial variable such as market value of fixed assets, gross value-added, total wage-bill, employment, amount of loan-outstanding, gross output, total cost of variable inputs and net-worth are winsorized at 1 and 99th percentile. Furthermore, the variables used in regressions $arpk$, $arpl$, k/l are winsorized at 1 and 99th percentile.

MSME Panel 2005-2007: The MSME dataset also provides retrospective information on output for the enterprises that survive upto 2007. This allows me to construct a balanced panel of MSMEs for the three-year period, 2005-2007. This panel information allows me to compute statistics such as average output growth, and volatility of output growth for each caste.

A.2 Economic Census 2005

The 5th Economic census in 2005 covered agricultural (excluding crop-production and plantation) and non-agricultural activities within the geographical boundary of India. The enterprise ownership across castes is measured with the Economic census of 2005 only for the noncorporate manufacturing enterprises. The caste of the private enterprise is identified with the caste of its owner.

The First column in Table A.1 is population shares, the second column is overall enterprise ownership by caste. In the third column, I report the share of one-employee enterprises owned by each caste. In the fourth column, I report the share of enterprises with hired workers (that hire workers outside their family members) owned by each caste. In the fifth column, I report the share of the total workforce employed by the enterprises of each caste. For instance, LC owns 16% of the overall enterprise, 16% of the one-employee enterprises, 11% of the enterprises with hired workers, and LC enterprises employ 11% of the total workforce in manufacturing.

In Table A.1b, I present the employment share distribution across enterprise size for the overall manufacturing sector and for each caste individually as well.

A.3 ASI-NSS 2006

Most commonly used firm-level dataset for India is a combination of Annual Survey of Industries (ASI, MOSPI 2005-2006) and the survey of the unregistered manufacturing sector from the National Sample Survey (NSS, MOSPI 2005-2006). The ASI focuses on the formal sector and covers enterprises employing 10 or more workers using electric power or employing 20 or more workers without electric power. The NSS, every five years, surveys a random sample of the population of manufacturing enterprises outside the sampling frame of the ASI. Hence, the firms in the NSS are decidedly smaller and mostly informal. This dataset has been used in [Hsieh and Klenow \(2009\)](#).

Table A.1: Economic Census 2005

(a) Panel A: Share of Population and Manufacturing Enterprises across castes

	(1)	(2)	(3)	(4)	(5)
Caste	Population	Enterprises	Enterprises with		
			One employee	Hired Workers	Employment
LC	31.2%	16.0%	16%	11%	11%
MC	39.3%	47.5%	51%	44%	40%
HC	29.5%	36.5%	33%	45%	49%

(b) Panel B: Cumulative Employment Share Distribution

Enterprises	p50	p75	p90	p95	p99
All	0.36	0.36	0.52	0.64	0.74
HC	0.24	0.32	0.48	0.56	0.68
MC	0.21	0.45	0.63	0.72	0.81
LC	0.23	0.54	0.66	0.78	0.86

Notes: The table reports descriptive statistics. LC, MC, and HC represents the low-caste, middle-caste, and high-caste, respectively.

Panel A. The population share in the first column are computed with IHDS data. The percentage of enterprises owned by each caste are reported in Column 2. The Column 3 shows the share of enterprises with one employee owned by each caste. In Column 4, Share of enterprises with hired labor for each caste is reported. In Column 5, the share of employees employed by the enterprises of different castes is reported (e.g., LC enterprises provide employment to 11% of the total work force).

Panel B. The share of employment distribution by caste. Each cell documents the percentage of employment employed by enterprises below a certain percentile (p) of size (employees).

Table A.2: Cumulative Income Distribution- IHDS

Households	p50	p75	p90	p95	p99
All	0.15	0.35	0.58	0.71	0.88
HC	0.14	0.37	0.57	0.70	0.88
MC	0.16	0.38	0.60	0.73	0.90
LC	0.19	0.40	0.63	0.75	0.91

Notes: The table reports descriptive statistics. LC, MC, and HC represents the low-caste, middle-caste, and high-caste households, respectively. Each cell documents the percentage of income owned by households below a certain percentile (p) of the income distribution.

A.4 Household Surveys 2004-05

Indian Human Development Survey is a nationally representative multi-topic panel survey of 42,152 households in 1,420 villages and 1,042 urban neighborhoods across India. IHDS 1 was conducted in 2005-2006, and IHDS 2 in 2011-2012. They include information on consumption, assets, wages, business-income, [Desai et al. \(2018\)](#). I define household income as income per capita. The resulting income distribution is provided in [Table A.2](#).

A.5 All India Debt and Investment Survey

All India Debt and Investment Survey (AIDIS) conducted in 2002-03. Wealth is computed as the total household assets net of the indebtedness. Household assets are defined as “physical assets like land, buildings, livestock, agricultural machinery and implements, non-farm business equipment, all transport equipment, durable household goods, and financial assets like dues receivable on loans

advanced in cash or in kind, shares in companies, and cooperative societies, banks, etc., national saving certificates and the like, deposits in companies, banks, post offices, and with individual. Debt is defined as cash loans payable. I use per capita wealth – household wealth divided by household size – as the measure of wealth.

Table A.3: Wealth Distribution: All India Debt and Investment Survey (AIDIS)

	p50	p75	p90	p95	p99
All Households					
All	0.07	0.23	0.45	0.59	0.81
HC	0.07	0.23	0.46	0.61	0.82
MC	0.09	0.27	0.50	0.64	0.85
LC	0.11	0.30	0.53	0.67	0.86
Entrepreneurs					
All	0.11	0.28	0.50	0.64	0.84
HC	0.10	0.28	0.50	0.64	0.84
MC	0.13	0.32	0.55	0.68	0.87
LC	0.15	0.35	0.58	0.71	0.88

Notes: The table reports descriptive statistics. LC, MC, and HC represents the low-caste, middle-caste, and high-caste, respectively. Each cell documents the percentage of wealth owned by households below a certain percentile (p) of the wealth distribution.

B Stylized Facts

B.1 Fact 1

The Table B.4 presents weighted average $\log(\text{ARPK})$, where weights are enterprise-level capital shares, in 2-digit sectors (in manufacturing). I have also reported the share of output for each caste in these sectors (computed in MSME dataset). Finally, the last column contains sectors' output share in the national manufacturing output (computed with ASI-NSS dataset). The weighted average $\log(\text{ARPK})$ of HC, MC, and LC enterprises in the aggregate economy is -1.80, -1.29, and -1.24 and unweighted average ARPK of HC, MC, and LC enterprises is -0.21, 0.02 and 0.25, respectively. This implies that weighted average ARPK of LC enterprise is 56% higher, whereas unweighted ARPK is 46% percent higher relative to HC enterprise.

Sampling Weights: Here, I provide the results for the main regression on $arpk$ with sampling weights. Table B.5 presents the regression estimates. The main results remain robust to the inclusion of sampling weights. Moreover, the $arpk$ for LC is 30% higher than HC (B.5), which is higher than what is reported in the main text. This is not surprising as the inclusion of sampling multipliers puts more weight on small enterprises (as these were sampled, see Section A) and these enterprises are more likely to be financially constrained.

ARPK in Service Sector: In the main text, I focused on the manufacturing sector, however, cross-caste $arpk$ differences remain high even in the service sector. This suggests that it is a within-sector phenomenon and, to a certain extent, the $arpk$ differences are driven by financial frictions. Table B.5 presents the regression estimates for service sector. The $arpk$ for LC and MC enterprises is 39% and

Table B.4: Sectoral Shares and ARPK across Castes

Sector	NIC Code	LC Share: output	$arpk_{LC}$	MC Share: output	$arpk_{MC}$	HC Share: output	$arpk_{HC}$	Sector Share: output
Manufacture of food products and beverages	15	0.016	-1.106	0.054	-1.445	0.106	-1.519	0.090
Manufacture of wearing apparel; dressing and dyeing of fur	18	0.013	-1.187	0.049	-0.943	0.052	-1.732	0.035
Manufacture of tobacco products	17	0.009	-1.446	0.040	-1.068	0.079	-1.745	0.081
Manufacture of furniture; manufacturing n.e.c.	36	0.008	-0.760	0.035	-0.815	0.036	-1.965	0.036
Manufacture of fabricated metal products, except machinery and equipment	28	0.006	-1.518	0.027	-1.398	0.057	-1.873	0.038
Manufacture of wood and of products of wood	20	0.006	-0.776	0.014	-1.188	0.017	-1.963	0.015
Manufacture of other non-metallic mineral products	26	0.003	-1.367	0.017	-1.293	0.043	-1.825	0.051
Manufacture of machinery and equipment n.e.c.	29	0.003	-1.511	0.014	-1.521	0.044	-1.963	0.057
Tanning and dressing of leather	19	0.002	-1.113	0.003	-1.864	0.008	-1.594	0.008
Manufacture of rubber and plastics products	25	0.002	-1.452	0.006	-1.858	0.024	-1.923	0.024
Manufacture of chemicals and chemical products	24	0.002	-1.577	0.009	-1.493	0.037	-1.953	0.133
Manufacture of basic metals	27	0.002	-1.499	0.006	-1.138	0.028	-1.742	0.113
Publishing, printing and reproduction of recorded media	22	0.002	-1.349	0.008	-1.333	0.014	-1.852	0.017
Manufacture of tobacco products	16	0.001	-0.274	0.004	-2.872	0.007	-1.098	0.020
Manufacture of electrical machinery and apparatus n.e.c.	31	0.001	-1.362	0.007	-0.957	0.017	-1.709	0.035
Manufacture of radio, television and communication equipment and apparatus	32	0.001	-0.979	0.002	-1.451	0.004	-2.291	0.015
Manufacture of other transport equipment	35	0.001	-1.292	0.003	-1.397	0.011	-1.731	0.025
Manufacture of medical, precision and optical instruments, watches and clocks	33	0.001	-1.709	0.003	-1.054	0.007	-2.032	0.008
Manufacture of paper and paper products	21	0.000	-1.482	0.003	-1.291	0.012	-1.860	0.014
Manufacture of motor vehicles, trailers and semi-trailers	34	0.000	-1.663	0.002	-0.988	0.005	-1.848	0.064
Manufacture of coke, refined petroleum products and nuclear fuel	23	0.000	-0.924	0.001	-0.746	0.002	-1.684	0.115
Manufacture of office, accounting and computing machinery	30	0.000	-1.380	0.000	-1.494	0.002	-2.135	0.007
Recycling	37	0.000	0.022	0.000	-0.872	0.000	-1.317	0.000
Total		0.08		0.31		0.61		1

Notes: The table presents weighted average ARPK (in logs), where weights are enterprise-level capital shares, in all the 2-digit sectors (in manufacturing). The table is sorted in terms of LC's output share (largest to smallest). The variables "LC share: output", "MC share: output" and "HC share: output" contain output share of LC, MC, and HC enterprises in that sector, respectively (computed in MSME dataset). The "sector share: output" contains sectors' output share in the national manufacturing output (computed with ASI-NSS dataset). The weighted average ARPK (in logs) of HC, MC, and LC enterprises in the aggregate economy is -1.80, -1.28, and -1.24, respectively. The Sampling multipliers are applied.

Table B.5: ARPK and ARPL across Castes

VARIABLES	(1) (2) (3) (4)				(5) (6) (7) (8)			
	Manufacturing				Services			
	Benchmark		With Sampling Weights		Benchmark		With Sampling Weights	
	log(ARPK)	log(ARPL)	log(ARPK)	log(ARPL)	log(ARPK)	log(ARPL)	log(ARPK)	log(ARPL)
M-Caste	0.13*** (0.02)	-0.05*** (0.01)	0.22*** (0.04)	0.09*** (0.03)	0.17*** (0.04)	0.11** (0.05)	0.09** (0.04)	0.10*** (0.04)
L-Caste	0.25*** (0.03)	-0.02 (0.02)	0.30*** (0.04)	0.14*** (0.05)	0.39*** (0.09)	0.20* (0.10)	0.19** (0.08)	0.15*** (0.05)
Constant	-0.67*** (0.01)	0.61*** (0.01)	-0.18*** (0.02)	0.59*** (0.02)	-0.51*** (0.03)	0.38*** (0.03)	0.03 (0.03)	0.76*** (0.03)
Observations	1,121,197	1,084,293	1,121,174	1,084,270	248,765	237,346	248,763	237,344
R-squared	0.51	0.31	0.38	0.37	0.54	0.42	0.46	0.54
REGION FE	YES	YES	YES	YES	YES	YES	YES	YES
NIC4 FE	YES	YES	YES	YES	YES	YES	YES	YES

Notes: Results from the enterprise level regression for manufacturing and service sector are presented separately. The variables ARPK and ARPL are the average products of capital and labor, respectively. M-caste is the dummy variable for the middle-caste enterprises. L-caste is the dummy variable for low-caste enterprises. All specification includes sector, district (586 in total), gender and religion of the owner fixed effects. The standard errors are in parentheses, clustered at caste, state and sector level. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

17% higher than HC enterprises. These values are a little bit higher than what is computed for the manufacturing sector.

Additional Controls: Further, I provide the robustness for the estimates for $arpk$. In the baseline, I control for fixed effects at 4 digit industry level. In Column 2 of Table B.6 I include fixed effect at 5 digit industry-level. Further, I provide robustness checks by including control variables that may be correlated with caste. Here, I include enterprise-level wages and mean schooling rates, land holdings, and dispersion in output growth that are computed at the sector-state-caste level. The $arpk$ differences across castes remain high (see Column 3 of Table B.6).

Table B.6: ARPK across Castes

VARIABLES	(1) log(ARPK)	(2) log(ARPK)	(3) log(ARPK)
M-Caste	0.13*** (0.02)	0.12*** (0.02)	0.09*** (0.02)
L-Caste	0.25*** (0.03)	0.23*** (0.03)	0.23*** (0.04)
wage			0.05** (0.02)
log(Years of Schooling.)			-0.04** (0.02)
log(land)			0.02** (0.01)
$sd(\Delta output)$			-0.21** (0.08)
Constant	-0.67*** (0.01)	-0.67*** (0.01)	-1.10*** (0.19)
Observations	1,121,197	1,121,174	910,794
R-squared	0.51	0.51	0.52
District FE	✓	✓	✓
NIC4 FE	✓		
NIC5 FE		✓	✓

Notes: Results from the enterprise level regression. M-caste is the dummy variable for the middle-caste enterprises and L-caste represents low-caste enterprises. All specification includes sector, state fixed effects, gender and religion of the owner fixed effects. $sd(\Delta output)$ is dispersion in growth rate, year of schooling, and land holdings are computed at sector, state and caste level. Robust standard errors are in parentheses, clustered at caste, region and sector levels. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

B.2 Fact 2

I provide the distribution of observations by caste in each size (output) bin in Table B.7. There are a substantial amount of enterprises for each caste in each of the bins. However, the LC enterprises are small so they are mostly present in the bottom three bins, whereas the majority of the HC enterprises are in the top three size bins.

Table B.7: Size Distribution of Enterprises by Caste in Different Size Bins

Caste	Size Bins					Total
	$\leq p10$	(p10-p50]	(p50-p75]	(p75-p90]	$\geq p90$	
HC	44,293	196,360	139,092	107,631	85,224	572,600
MC	50,252	198,407	117,529	50,808	21,982	438,978
LC	17,602	55,098	22,454	9,752	4,935	109,841
Total	112,147	449,865	279,075	168,191	112,141	1,121,419

Notes: The table reports the number of observation by caste for size bin (size is defined by enterprise output). The size bins are defined as: (i) p10: enterprises below 10th percentile, (ii) p10-p50: enterprises between 10th and 50th percentile (iii) p50-p75: enterprises between 50th and 75th percentile, (iv) p75-p90: enterprises between 75th and 90th percentile, and (v) p90: enterprises above 90th percentile. The standard are clustered at caste, state and sector level.

I re-estimate the coefficients reported in Figure 1 in the main text with sampling multipliers. Going from enterprises in the bottom decile to enterprises in the top decile of the size (output) distribution, $arpk$ differences between LC and HC enterprises decline by 20 p.p. This is smaller than what is document in the main text but the overall trend seems to be in the right direction. The $arpk$ difference between MC and HC enterprises decline by 12 p.p. over enterprises size (See Figure B.1).

Table B.8: Dynamics of ARPK over Age for LC & MC Entrepreneurs

	LC	MC
Young enterprises (Age < 10)	+25.6%	+13.1%
Large enterprises (Age > 30 & Age < 40)	+22.1%	+10.8%

Notes: The measures of *arpk* are computed for MSMEs with respect to HC entrepreneurs, while controlling for sectoral, regional, gender and religion of the owner fixed effects. Young enterprises are aged less than 10 years and old enterprises are aged between 30 and 40 years. LC are low castes and MC are middle castes.

Further, age of a enterprise is positively correlated with size and thus it may also have implication for *arpk* dynamics. Therefore, I compare the young (age less than ten years) and old enterprises (age between thirty and forty years) and report the cross-caste differences in *arpk*. The *arpk* is highest for young LC entrepreneurs, +26% relative to that of HC establishments, and it declines by 4 p.p. over age and it stands at 22% for old enterprises. Similarly, over age, a decline of 2 p.p. in *arpk* (relative to HC enterprises) is estimated for MC enterprises.

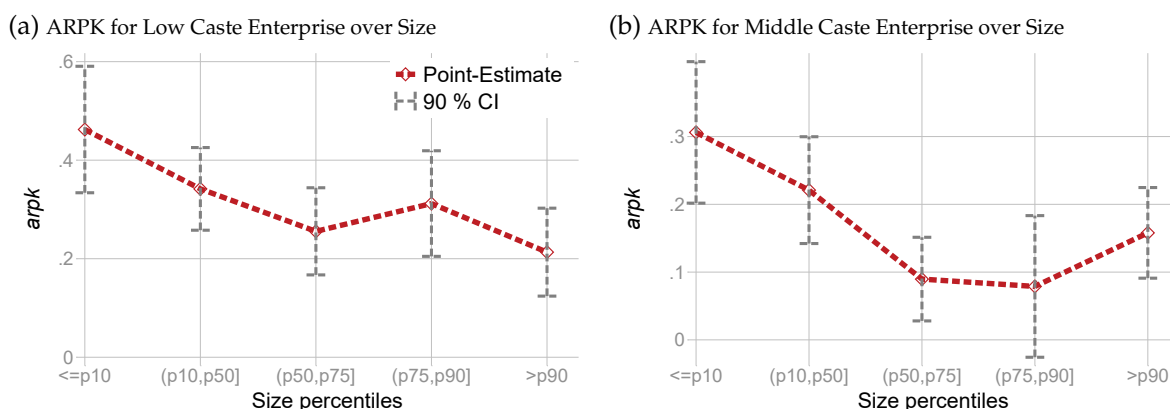


Figure B.1: ARPK over Enterprise Size

Notes: This figure plots the evolution of *arpk* over enterprise size (output). The size bins are defined as: (i) p10: enterprises below 10th percentile, (ii) p10-p50: enterprises between 10th and 50th percentile (iii) p50-p75: enterprises between 50th and 75th percentile, (iv) p75-p90: enterprises between 75th and 90th percentile, and (v) p90: enterprises above 90th percentile. The standard are clustered at caste, state and sector level.

B.3 Fact 3

In Table B.9, I re-estimate the coefficients for Fact 3 with sampling weights. The results are qualitatively similar. For instance, the *arpk* for LC and MC enterprises is relative to HC enterprises on average but these differences decline as the regions become more financially developed. Similarly, the capital-labor ratio for LC enterprises increases more rapidly relative to HC enterprises.

There may exist several factors other than the credit-to-output ratio that differ across regions. For instance, dispersion in capital-to-output ratio varies substantially across regions. I compute the standard deviation in the capital-output ratio within an industry and then compute a weighted average across industries within a state, where weights are industry output shares (see Table B.10). In Table B.11, I report estimates of regression model used in Fact 3 with a set of regions with comparable dis-

Table B.9: Regional Financial Development and Capital Allocation

	log(ARPK)	log(ARPK)	log(K/L)	log(K/L)	Profit Rate	Profit Rate
M-Caste	0.22*** (0.04)	0.36*** (0.10)	-0.27*** (0.04)	-0.44*** (0.11)	0.10*** (0.02)	0.23*** (0.05)
L-Caste	0.30*** (0.04)	0.57*** (0.10)	-0.45*** (0.05)	-0.92*** (0.11)	0.14*** (0.03)	0.34*** (0.07)
FD_s		-0.16* (0.09)		0.03 (0.12)		-0.04 (0.05)
M-Caste \times FD_s		-0.29* (0.15)		0.20 (0.16)		-0.37*** (0.11)
L-Caste \times FD_s		-0.45** (0.18)		0.66*** (0.22)		-0.41*** (0.12)
Constant	-0.18*** (0.02)	-0.12* (0.07)	10.51*** (0.02)	10.57*** (0.08)	0.09*** (0.01)	0.12*** (0.03)
Observations	1,119,485	1,119,491	1,119,485	1,119,491	1,119,485	1,119,491
R-squared	0.38	0.10	0.47	0.18	0.29	0.06
District FE	✓	-	✓	-	✓	-
NIC4 FE	✓	✓	✓	✓	✓	✓

Notes: Results from the firm-level regression used in Fact 3 in the main text. The Profit rate is IHS transformed as shown in column headings. Fd is an index of financial development across states. The variables ARPK, K/L, and Profit rate are the average products of capital, capital-labor ratio and profit rate. M-caste is the dummy variable for the middle-caste and L-caste is the dummy variable for low-caste enterprises. All specification includes sector (211 in total), gender and religion of the owner fixed effects. Robust standard errors are in parentheses, clustered at the caste, region and sector level. Sampling weights are applied. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

persion in capital-to-output ratio. The restricted sample includes regions where the dispersion in the capital-to-output ratio is between 1.5 and 4.

Further, I report the caste composition of the population for each region in Table B.12. I control for LC's population share in regression used in Fact 3 and report estimates in Table B.13. Further, in order to get rid of the regions which are dominated by households of a certain caste, I restrict the sample to a set of regions where LC population share varies between 20% to 40%. I re-run the regression specification for Fact 3 and found qualitatively similar results.

Further, I use other proxies for regional financial development such as: a) Commercial banks per capita (in 10^{-3}) is the number of commercial bank branches per capita in a region, b) Rural banks per capita (in 10^{-3}) is rural bank branches in a region, c) All banks per capita (in 10^{-3}) is the sum of commercial and rural banks per capita in a region, d) % of HH with Bank account is the percentage of households with an active bank account, and e) % of HH with a loan is the percentage of households who borrowed money from banks in the past five years. I find qualitatively similar results for all proxies of regional financial development. I provide the results in Table B.14, B.15 and B.16.

Table B.10: Regional heterogeneity in dispersion in capital-output ratio

Region	Share of output	Share of firms	Population Share	SD(K/Y)
Tamil Nadu	0.121	0.120	0.057	2.643
West Bengal	0.118	0.122	0.075	1.541
Gujarat	0.117	0.076	0.049	3.873
Uttar Pradesh	0.101	0.123	0.156	1.602
Maharashtra	0.095	0.064	0.090	2.898
Andhra Pradesh	0.070	0.062	0.071	1.781
Karnataka	0.061	0.081	0.052	2.312
Kerala	0.055	0.063	0.028	2.632
Punjab	0.044	0.022	0.022	2.768
Haryana	0.036	0.018	0.019	2.425
Rajasthan	0.032	0.032	0.057	2.445
Odisha	0.027	0.053	0.035	1.459
Madhya Pradesh	0.020	0.040	0.053	1.310
Bihar	0.020	0.041	0.073	1.269
Jharkhand	0.014	0.025	0.039	0.722
Assam	0.011	0.010	0.025	1.616
Uttarakhand	0.008	0.011	0.016	1.907
Himachal Pradesh	0.008	0.010	0.006	2.079
Jammu and Kashmir	0.008	0.007	0.012	2.131
Delhi	0.007	0.003	0.017	1.838
Chhattisgarh	0.005	0.008	0.029	1.856
Nagaland	0.004	0.001	0.002	1.741
Goa	0.003	0.002	0.003	1.786
Puducherry	0.002	0.000	0.002	2.475
Manipur	0.002	0.001	0.002	1.047
Chandigarh	0.001	0.000	0.001	2.474
Arunachal Pradesh	0.001	0.001	0.001	0.693
Tripura	0.001	0.001	0.003	1.869
Meghalaya	0.001	0.001	0.002	1.093
Mizoram	0.001	0.000	0.001	1.840
Sikkim	0.000	0.000	0.001	4.057

Notes: This table present the regional share of output, firms, population, and the standard deviation in capital-output ratio. I compute the standard deviation within an industry and then compute a weighted average across industries within a state, where weights are industry output shares. SD(K/Y) represents the standard deviation in capital-output ratio within a region. Sampling weights are applied.

Table B.11: Regional Financial Development and Capital Allocation

	(1)	(2)	(3)	(4)	(5)	(6)
	Restricted Sample					
	log(ARPK)	log(ARPK)	log(K/L)	log(K/L)	Profit Rate)	Profit Rate
M-caste	0.11*** (0.02)	1.01*** (0.23)	-0.23*** (0.02)	-0.97*** (0.20)	0.02* (0.01)	0.67*** (0.16)
L-caste	0.24*** (0.03)	1.22*** (0.24)	-0.47*** (0.03)	-1.33*** (0.20)	0.06*** (0.02)	0.68*** (0.18)
FD_s		1.10*** (0.21)		-0.55*** (0.18)		0.84*** (0.15)
M-caste \times FD_s		-0.90*** (0.26)		0.56** (0.24)		-0.72*** (0.18)
L-caste \times FD_s		-1.00*** (0.27)		0.76*** (0.24)		-0.68*** (0.19)
Constant	-0.80*** (0.01)	-1.55*** (0.20)	11.51*** (0.02)	12.00*** (0.17)	-0.11*** (0.01)	-0.66*** (0.15)
Observations	997,397	997,401	997,397	997,401	997,397	997,401
R-squared	0.49	0.13	0.53	0.22	0.46	0.10
District FE	✓	-	✓	-	✓	-
NIC4 FE	✓	✓	✓	✓	✓	✓

Notes: Results from the firm-level regression used in Fact 3 in the main text. The Profit rate is IHS transformed as shown in column headings. M-caste is the dummy variable for the middle-caste and L-caste is the dummy variable for low-caste enterprises. Fd is an index of financial development across states. The variables ARPK, K/L, and Profit rate are the average products of capital, capital-labor ratio and profit rate. $SD\left(\frac{K}{Y}\right)_s$ represents the standard deviation in capital-output ratio within a region. The restricted sample includes regions where $SD\left(\frac{K}{Y}\right)_s$ is between 1.5 and 4. All specification includes sector (211 in total), gender and religion of the owner fixed effects. Robust standard errors are in parentheses, clustered at the caste, region and sector level. Sampling weights are applied. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.12: Population Share across castes

Region	Population Share			
	LC	MC	HC	Overall
Uttar Pradesh	0.237	0.535	0.228	0.156
Maharashtra	0.257	0.336	0.407	0.090
West Bengal	0.404	0.094	0.501	0.075
Bihar	0.228	0.592	0.180	0.073
Andhra Pradesh	0.286	0.544	0.170	0.071
Tamil Nadu	0.310	0.623	0.068	0.057
Rajasthan	0.300	0.493	0.206	0.057
Madhya Pradesh	0.340	0.447	0.213	0.053
Karnataka	0.278	0.504	0.218	0.052
Gujarat	0.230	0.457	0.312	0.049
Jharkhand	0.448	0.351	0.202	0.039
Orissa	0.385	0.437	0.178	0.035
Chhatisgarh	0.410	0.487	0.104	0.029
Kerala	0.105	0.568	0.327	0.028
Assam	0.369	0.106	0.525	0.025
Punjab	0.391	0.225	0.383	0.022
Haryana	0.278	0.357	0.365	0.019
Delhi	0.321	0.300	0.379	0.017
Uttaranchal	0.409	0.249	0.342	0.016
Jammu & Kashmir	0.115	0.157	0.728	0.012
Himachal Pradesh	0.326	0.089	0.585	0.006
Tripura	0.460	0.211	0.329	0.003
Goa	0.138	0.310	0.552	0.003
Meghalaya	0.845	0	0.155	0.002
Manipur	0	0.063	0.937	0.002
Nagaland	0.883	0	0.117	0.002
Pondicherry	0.041	0.923	0.035	0.002
Chandigarh	0.134	0.254	0.612	0.001
Arunachal Pradesh	0.911	0.028	0.061	0.001
Dadra+Nagar Haveli	0.388	0.245	0.367	0.001
Sikkim	0.225	0.414	0.361	0.001
Daman & Diu	0.434	0.470	0.096	0.000

Notes: The Table presents population share by caste within a region and the population share of each region in the country using IHDS data. HC, MC and LC represents high castes, middle castes and low castes, respectively. Few regions (small) are missing due to lack of enterprises in the sample. Sampling weights are applied

Table B.13: Regional Financial Development and Capital Allocation

	Full Sample						Restricted Sample					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	log(ARPK)	log(ARPK)	log(K/L)	log(K/L)	Profit Rate	Profit Rate	log(ARPK)	log(ARPK)	log(K/L)	log(K/L)	Profit Rate	Profit Rate
M-Caste	0.14*** (0.02)	1.35*** (0.21)	-0.30*** (0.02)	-1.41*** (0.19)	0.03*** (0.01)	0.77*** (0.15)	0.18*** (0.03)	2.14*** (0.29)	-0.35*** (0.03)	-2.16*** (0.24)	0.06*** (0.02)	1.26*** (0.20)
L-Caste	0.25*** (0.03)	1.37*** (0.22)	-0.49*** (0.03)	-1.52*** (0.19)	0.05*** (0.02)	0.72*** (0.16)	0.28*** (0.03)	1.55*** (0.28)	-0.51*** (0.04)	-1.70*** (0.24)	0.07*** (0.02)	0.82*** (0.19)
FD_s		0.83*** (0.20)		-0.31* (0.17)		0.71*** (0.14)		1.16*** (0.22)		-0.59*** (0.18)		0.92*** (0.15)
M-Caste \times FD_s		-1.36*** (0.23)		1.05*** (0.21)		-0.80*** (0.15)		-1.69*** (0.24)		1.42*** (0.22)		-0.97*** (0.16)
L-Caste \times FD_s		-1.26*** (0.24)		1.05*** (0.23)		-0.76*** (0.16)		-1.35*** (0.30)		1.17*** (0.27)		-0.79*** (0.20)
Population Share	-0.04 (0.05)	-0.37** (0.17)	0.18** (0.07)	0.64*** (0.17)	-0.05** (0.02)	-0.33*** (0.11)	-0.13* (0.07)	-1.91*** (0.35)	0.29*** (0.10)	1.95*** (0.33)	-0.12*** (0.04)	-1.36*** (0.23)
Constant	-0.66*** (0.02)	-1.19*** (0.19)	11.32*** (0.03)	11.57*** (0.16)	-0.04*** (0.01)	-0.43*** (0.13)	-0.72*** (0.02)	-1.10*** (0.19)	11.36*** (0.03)	11.50*** (0.16)	-0.07*** (0.01)	-0.36*** (0.13)
Observations	1,118,427	1,118,433	1,118,427	1,118,433	1,118,427	1,118,433	897,456	897,458	897,456	897,458	897,456	897,458
R-squared	0.51	0.13	0.55	0.23	0.46	0.09	0.55	0.19	0.59	0.29	0.50	0.14
District FE	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
NIC4 FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Results from the firm-level regression used in Fact 3 in the main text. The Profit rate is IHS transformed as shown in column headings. M-caste is the dummy variable for the middle-caste and L-caste is the dummy variable for low-caste enterprises. Fd is an index of financial development across states. The variables ARPK, K/L, and Profit rate are the average products of capital, capital-labor ratio and profit rate. Population share represents the regional population share of each caste. The restricted sample includes regions where LC population share is between 20% to 40%. All specification includes sector (211 in total), gender and religion of the owner fixed effects. Robust standard errors are in parentheses, clustered at the caste, region and sector level. Sampling weights are applied. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.14: Regional Financial Development and ARPK

VARIABLES	(1) log(ARPK)	(2) log(ARPK)	(3) log(ARPK)	(4) log(ARPK)	(5) log(ARPK)	(6) log(ARPK)
M-Caste	0.13*** (0.02)	1.33*** (0.21)	1.71*** (0.30)	1.71*** (0.33)	0.65*** (0.25)	2.11*** (0.28)
L-Caste	0.25*** (0.03)	0.95*** (0.22)	1.81*** (0.30)	1.06*** (0.32)	1.05*** (0.24)	1.77*** (0.29)
Commercial banks per capita (10^{-3})		-3.28** (1.33)				
M-Caste \times Commercial banks per capita (10^{-3})		-8.33*** (1.75)				
L-Caste \times Commercial banks per capita (10^{-3})		-2.07 (1.97)				
Rural banks per capita (10^{-3})			121.49*** (14.27)			
M-Caste \times Rural banks per capita (10^{-3})			-81.15*** (16.14)			
L-Caste \times Rural banks per capita (10^{-3})			-86.87*** (15.76)			
All Banks per capita (10^{-3})				2.14 (1.94)		
M-Caste \times All banks per capita (10^{-3})				-11.36*** (2.53)		
L-Caste \times All banks per capita (10^{-3})				-2.80 (2.41)		
% of HH with Bank account					-1.00*** (0.37)	
M-Caste \times % of HH with Bank account					-0.21 (0.64)	
L-Caste \times % of HH with Bank account					-1.01* (0.58)	
% of HH with loan						8.58*** (1.57)
M-Caste \times % of HH with loan						-12.20*** (1.69)
L-Caste \times % of HH with loan						-8.48*** (1.70)
Constant	-0.68*** (0.01)	-0.66*** (0.17)	-2.47*** (0.26)	-1.12*** (0.26)	-0.62*** (0.15)	-2.05*** (0.25)
Observations	1,119,508	1,119,514	1,114,209	1,114,209	1,119,514	1,119,514
R-squared	0.51	0.13	0.22	0.13	0.12	0.16
District FE	Yes	No	No	No	No	No
NIC4 FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Results from the firm-level regression used in Fact 3 in the main text. Dependent variable is average products of capital ARPK. M-caste is the dummy variable for the middle-caste and L-caste is the dummy variable for low-caste enterprises. Commercial banks per capita (in 10^{-3}) is number of commercial bank branches per capita in a region. Rural banks per capita (in 10^{-3}) is rural bank branches in a region. All banks per capita (in 10^{-3}) is sum of commercial and rural banks per capita in a region. % of HH with Bank account is percentage of households with a working bank account and % of HH with loan is percentage of households who borrowed money from banks in the past five years. All specification includes sector (211 in total), gender and religion of the owner fixed effects. Robust standard errors are in parentheses, clustered at the caste, region and sector level. Sampling weights are applied. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.15: Regional Financial Development and Capital-labor ratio

VARIABLES	(1) log(K/L)	(2) log(K/L)	(3) log(K/L)	(4) log(K/L)	(5) log(K/L)	(6) log(K/L)
M-Caste	-0.26*** (0.02)	-1.51*** (0.20)	-1.64*** (0.24)	-1.81*** (0.30)	-0.74*** (0.23)	-2.23*** (0.24)
L-Caste	-0.49*** (0.03)	-1.13*** (0.21)	-1.92*** (0.25)	-1.15*** (0.29)	-1.36*** (0.23)	-1.76*** (0.24)
Commercial banks per capita (10^{-3})		5.37*** (1.35)				
M-Caste \times Commercial banks per capita (10^{-3})		8.93*** (1.85)				
L-Caste \times Commercial banks per capita (10^{-3})		1.46 (2.07)				
Rural banks per capita (10^{-3})			-112.77*** (11.45)			
M-Caste \times Rural banks per capita (10^{-3})			65.72*** (13.66)			
L-Caste \times Rural banks per capita (10^{-3})			77.62*** (13.31)			
All banks per capita (10^{-3})				0.39 (1.89)		
M-Caste \times All banks per capita (10^{-3})				11.09*** (2.55)		
L-Caste \times All banks per capita (10^{-3})				1.35 (2.37)		
% of HH with Bank account					1.75*** (0.31)	
M-Caste \times % of HH with Bank account					0.20 (0.65)	
L-Caste \times % of HH with Bank account					1.36** (0.59)	
% of HH with loan						-7.39*** (1.27)
M-Caste \times % of HH with loan						11.86*** (1.49)
L-Caste \times % of HH with loan						6.43*** (1.46)
Constant	11.38*** (0.02)	11.19*** (0.15)	13.05*** (0.21)	11.59*** (0.23)	11.09*** (0.13)	12.45*** (0.21)
Observations	1,119,508	1,119,514	1,114,209	1,114,209	1,119,514	1,119,514
R-squared	0.55	0.24	0.31	0.23	0.24	0.26
District FE	Yes	No	No	No	No	No
NIC4 FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Results from the firm-level regression used in Fact 3 in the main text. Dependent variable is capital-labor ratio K/L . M-caste is the dummy variable for the middle-caste and L-caste is the dummy variable for low-caste enterprises. Commercial banks per capita (in 10^{-3}) is number of commercial bank branches per capita in a region. Rural banks per capita (in 10^{-3}) is rural bank branches in a region. All banks per capita (in 10^{-3}) is sum of commercial and rural banks per capita in a region. % of HH with Bank account is percentage of households with a working bank account and % of HH with loan is percentage of households who borrowed money from banks in the past five years. All specification includes sector (211 in total), gender and religion of the owner fixed effects. Robust standard errors are in parentheses, clustered at the caste, region and sector level. Sampling weights are applied. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table B.16: Regional Financial Development and Profit Rate

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)
	Profit Rate	Profit Rate	Profit Rate	Profit Rate	Profit Rate	Profit Rate
M-Caste	0.02** (0.01)	0.62*** (0.15)	1.00*** (0.22)	0.88*** (0.23)	0.35** (0.16)	1.16*** (0.20)
L-Caste	0.05*** (0.02)	0.44*** (0.15)	1.08*** (0.22)	0.63*** (0.21)	0.49*** (0.15)	1.00*** (0.20)
Commercial banks per capita (10^{-3})		-0.66 (0.90)				
M-Caste \times Commercial banks per capita (10^{-3})		-3.51*** (1.20)				
L-Caste \times Commercial banks per capita (10^{-3})		-0.95 (1.24)				
Rural banks per capita (10^{-3})			74.66*** (10.50)			
M-Caste \times Rural banks per capita (10^{-3})			-50.72*** (11.65)			
L-Caste \times Rural banks per capita (10^{-3})			-59.25*** (11.54)			
All banks per capita (10^{-3})				3.20** (1.32)		
M-Caste \times All banks per capita (10^{-3})				-5.84*** (1.71)		
L-Caste \times All banks per capita (10^{-3})				-2.80* (1.52)		
% of HH with Bank account					-0.34 (0.25)	
M-Caste \times % of HH with Bank account					-0.12 (0.38)	
L-Caste \times % of HH with Bank account					-0.49 (0.35)	
% of HH with loan						6.55*** (1.11)
M-Caste \times % of HH with loan						-7.25*** (1.16)
L-Caste \times % of HH with loan						-5.57*** (1.13)
Constant	-0.05*** (0.01)	-0.15 (0.12)	-1.15*** (0.20)	-0.50*** (0.18)	-0.10 (0.10)	-0.93*** (0.19)
Observations	1,119,508	1,119,514	1,114,209	1,114,209	1,119,514	1,119,514
R-squared	0.46	0.08	0.15	0.08	0.08	0.12
District FE	Yes	No	No	No	No	No
NIC4 FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Results from the firm-level regression used in Fact 3 in the main text. Dependent variables Profit rate is IHS transformed. M-caste is the dummy variable for the middle-caste and L-caste is the dummy variable for low-caste enterprises. Commercial banks per capita (in 10^{-3}) is number of commercial bank branches per capita in a region. Rural banks per capita (in 10^{-3}) is rural bank branches in a region. All banks per capita (in 10^{-3}) is sum of commercial and rural banks per capita in a region. % of HH with Bank account is percentage of households with a working bank account and % of HH with loan is percentage of households who borrowed money from banks in the past five years. All specification includes sector (211 in total), gender and religion of the owner fixed effects. Robust standard errors are in parentheses, clustered at the caste, region and sector level. Sampling weights are applied. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

C Model

Proposition 1: Under perfect credit Benchmark, the employment size distribution follows a power law;

$$Pr[L > \ell] = \left(\frac{\ell(\hat{z})}{\ell} \right)^{v\eta}, \quad \ell \geq \ell(\hat{z}), \quad (1)$$

where, $\ell(\hat{z})$ is labor demand of the marginal entrant \hat{z} is the threshold productivity to become and entrepreneur.

Proof: The productivity of entrepreneurs, z , follows a Pareto distribution with scale parameter η . Labor demand of an entrepreneur of productivity z can be written as (in terms of labor demand of marginal entrepreneur),

$$l(z) = \left(\frac{z}{\hat{z}} \right)^{\frac{1}{v}} \ell(\hat{z}). \quad (2)$$

$$Pr[\tilde{\ell} > \ell] = Pr \left[z > \left(\frac{\ell(\hat{z})}{\ell} \right)^{-v} \hat{z} \mid z > \hat{z} \right] = \left(\frac{\ell(\hat{z})}{\ell} \right)^{v\eta}. \quad (3)$$

C.1 Income and Wealth Inequality

In this section, I report the income and wealth inequality estimates in the model. There is huge cross-caste inequality in income and wealth in the data (see Table C.17). Low castes earn half of what high castes earn on average, whereas low castes have one-third of the wealth of an average high caste. The model also produces results that are qualitatively similar in these two aspects, however, the model underestimates the cross-caste inequality. This may suggest that caste-specific borrowing constraints do make low castes poorer on average but are not enough to match the cross-caste disparities in income and wealth as documented in the data.

In Figure C.2, I compare the distribution of income for households within each caste. The model does well in capturing the upper tail of the income distribution for each caste. In particular, income distribution is more fat-tailed for HC households relative to LC households both in the data and in the model.

In the counterfactual experiment, cross-caste wealth inequality declines as LC households become richer on average relative to the benchmark economy (wealth increases by 16.4% for LC on average). Moreover, when LC enterprises are allowed to borrow more, the wealth inequality within LC households increases as the upper tail of the distribution is able to create more wealth (see Table C.18). The cross-caste income inequality decreases as the average LC household is richer in the counterfactual economy (wealth increases by 8.8% for LC on average).

Table C.17: Income and Wealth Inequality across Castes

	Income		Wealth		Counterfactual Gains	
	Model	Data	Model	Data	Income	Wealth
HC	1.00	1.00	1.00	1.00	1.6%	-5.9%
MC	0.97	0.76	0.92	0.53	6.6%	6.5%
LC	0.91	0.51	0.74	0.28	8.8%	16.4%

Notes: The table compares income and wealth disparities across castes. The income is computed as average income-per-capita for households in IHDS 2004-05 (Section A.4) and then normalized to its value for HC households. The wealth is networth-per-capita computed in AIDIS 2002 (Section A.5) and normalized to its value for HC households. The normalization helps to gauge caste disparities. The counterfactual gains represent the economy where borrowing constraints are equalized. The gains in income and wealth for each caste are computed relative to their respective values in the benchmark economy.

Table C.18: Cumulative Income and Wealth Distribution-Model

Enterprises	Income Distribution: Households					Asset Distribution: Entrepreneurs		
	p50 (%)	p75 (%)	p90 (%)	p95 (%)	p99 (%)	p90 (%)	p95 (%)	p95 (%)
Panel a. Benchmark Economy								
HC	35.3	53.1	65.2	72.4	85.7	36.5	52.3	81.2
MC	36.5	54.7	66.9	73.7	86.2	41.1	56.4	81.1
LC	44.5	57.8	69.4	74.7	85.9	41.1	57.6	85.5
Panel b. Counterfactual Economy								
HC	36.9	55.4	67.8	74.5	87.0	37.0	53.2	81.0
MC	36.5	55.8	67.3	74.4	87.0	35.7	51.0	79.0
LC	40.1	56.5	68.4	74.8	86.6	38.4	55.0	83.0

Notes: This table presents cumulative income and wealth distribution. Each cell documents the percentage of income/assets owned by households below a certain percentile (p). Panel (a) presents income distribution for households and wealth distribution for entrepreneurs for each caste in the benchmark economy. Panel (b) presents income distribution for households and wealth distribution for entrepreneurs for each caste in the counterfactual economy.

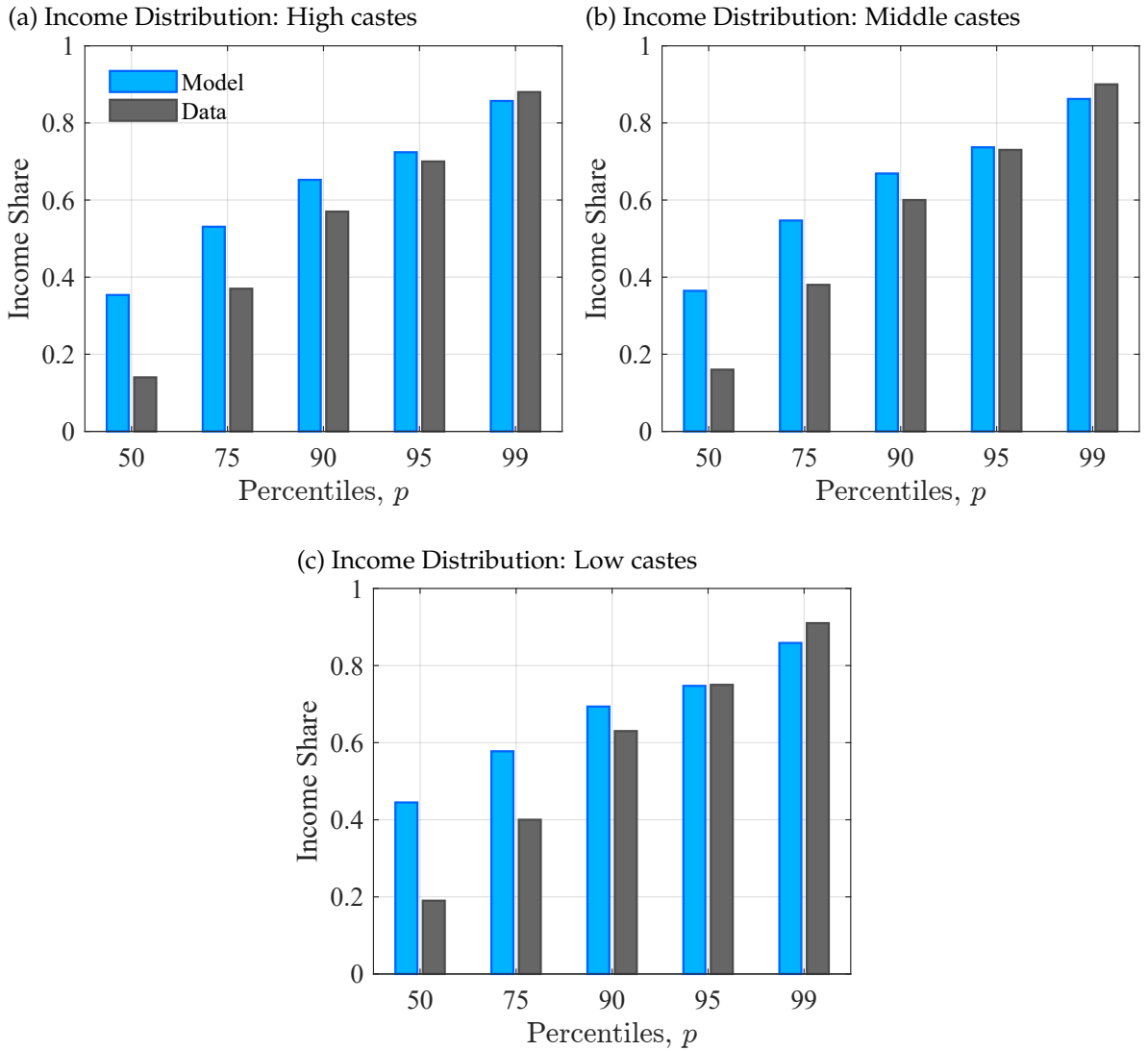


Figure C.2: Income Distribution across Castes

Notes: Each bar represents the percentage of income owned by households below a certain percentile (p) of the income distribution. Grey bars represent data points taken from IHDS, and blue bars represent model counterparts.

References

- Sonalde Desai, Reeve Vanneman, and New Delhi National Council of Applied Economic Research. *India Human Development Survey (IHDS), 2005*. Inter-university Consortium for Political and Social Research [distributor], 2018-08-08. <https://doi.org/10.3886/ICPSR22626.v12>, 2018.
- Chang-Tai Hsieh and Peter J. Klenow. Misallocation and manufacturing TFP in China and India. *The Quarterly Journal of Economics*, CXXIV(November), 2009.
- MSME. *Fourth All India Census of Micro, Small and Medium Enterprises, 2006*. Ministry of Micro, Small and Medium Enterprises, 2009. URL https://www.dcmsme.gov.in/ito_msme/censuses.htm.