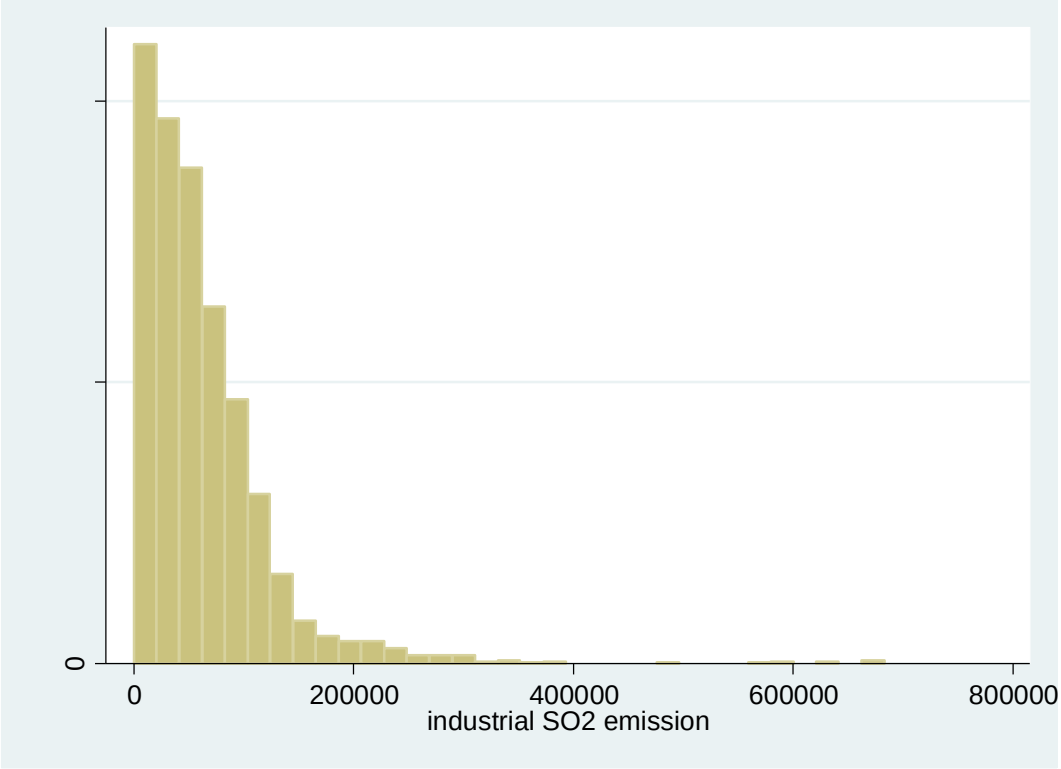
Data appendix

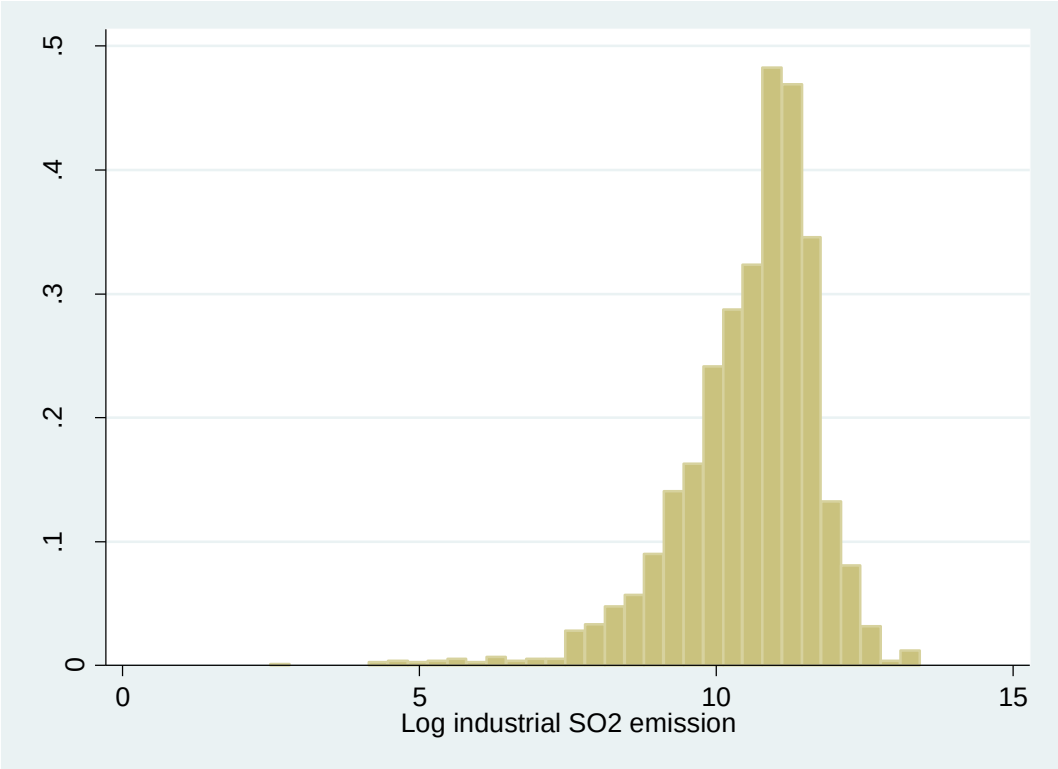
Data source: All data come from China City Statistical Yearbook 2004-2011

Quantitative variables:

1) Total industrial SO2 emissions
Variable name: so2_emi
Missing observations: 22/2296
Description: Total emission of industrial sulfur dioxide (ton) of a city
Descriptive statistics:
Mean: 63014.15
Median: 49708.50
Standard deviation: 63692.45
Minimum: 12
Maximum: 683162

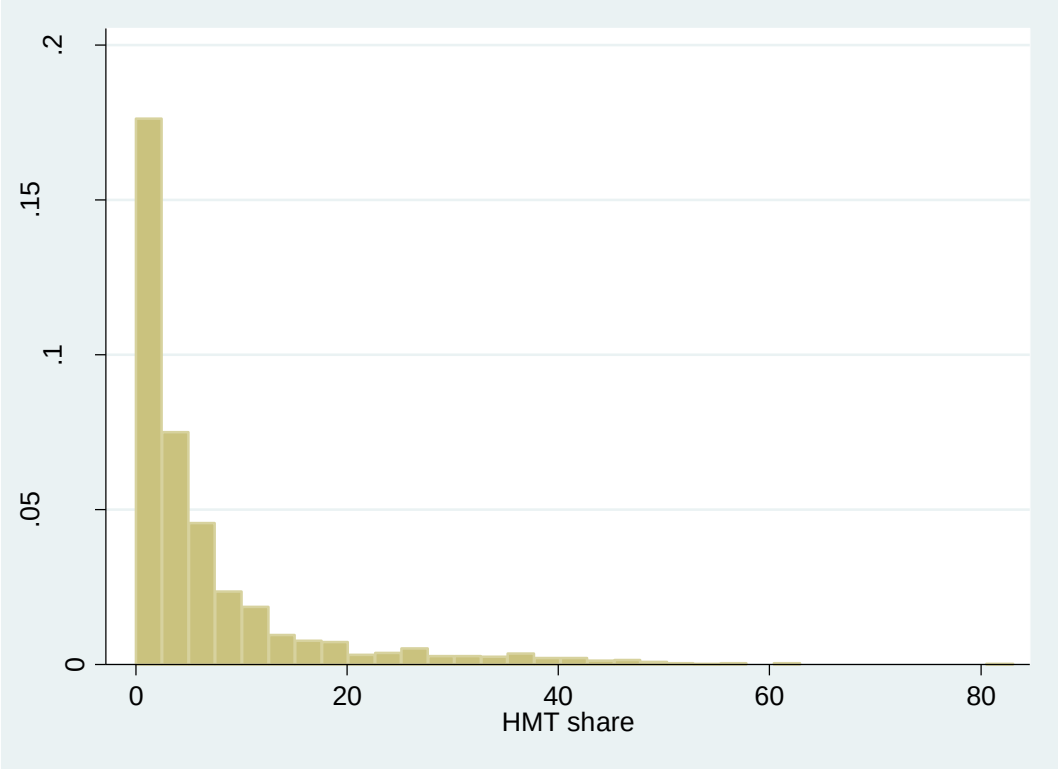


Histogram of log transformed form of the variable:



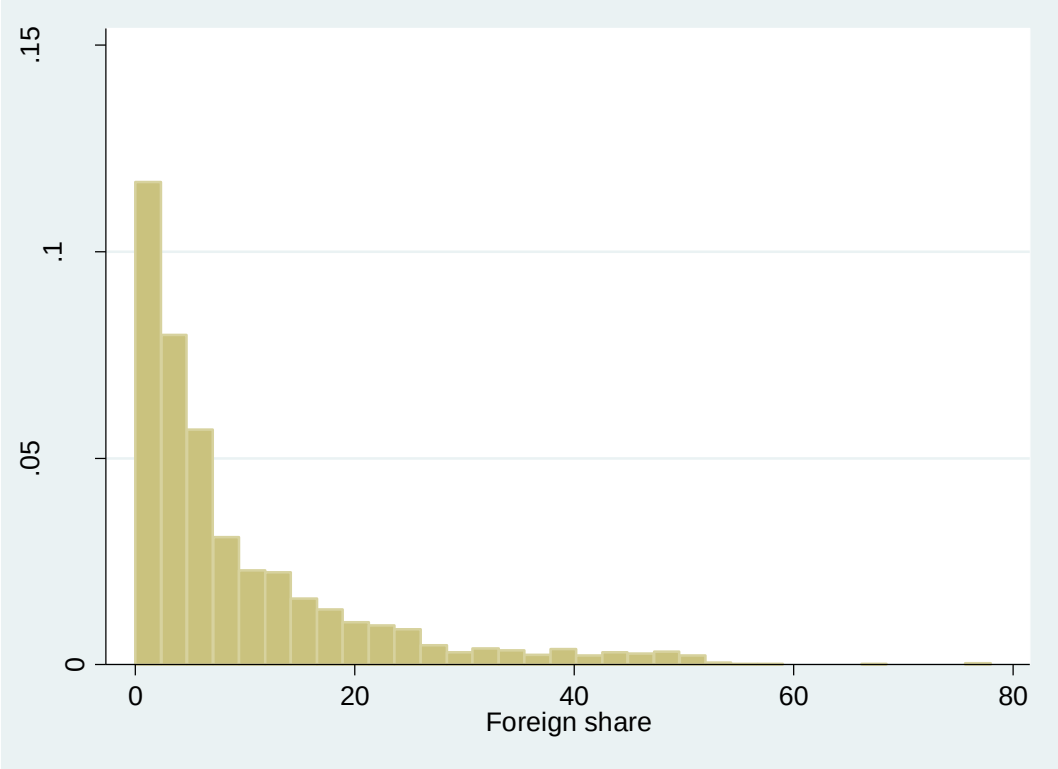
2) Share of total industrial output by Hong Kong, Macau, and Taiwan (HMT) firms

Variable name: ind_val_hmt_per
Missing observations: 152/2296
Description: Percentage (%) of combined industrial output by Hong Kong, Macau, and Taiwan (HMT) firms in total industrial output
Descriptive statistics:
Mean: 6.77
Median: 3.09
Standard deviation: 9.61
Minimum: 0
Maximum: 83.02312



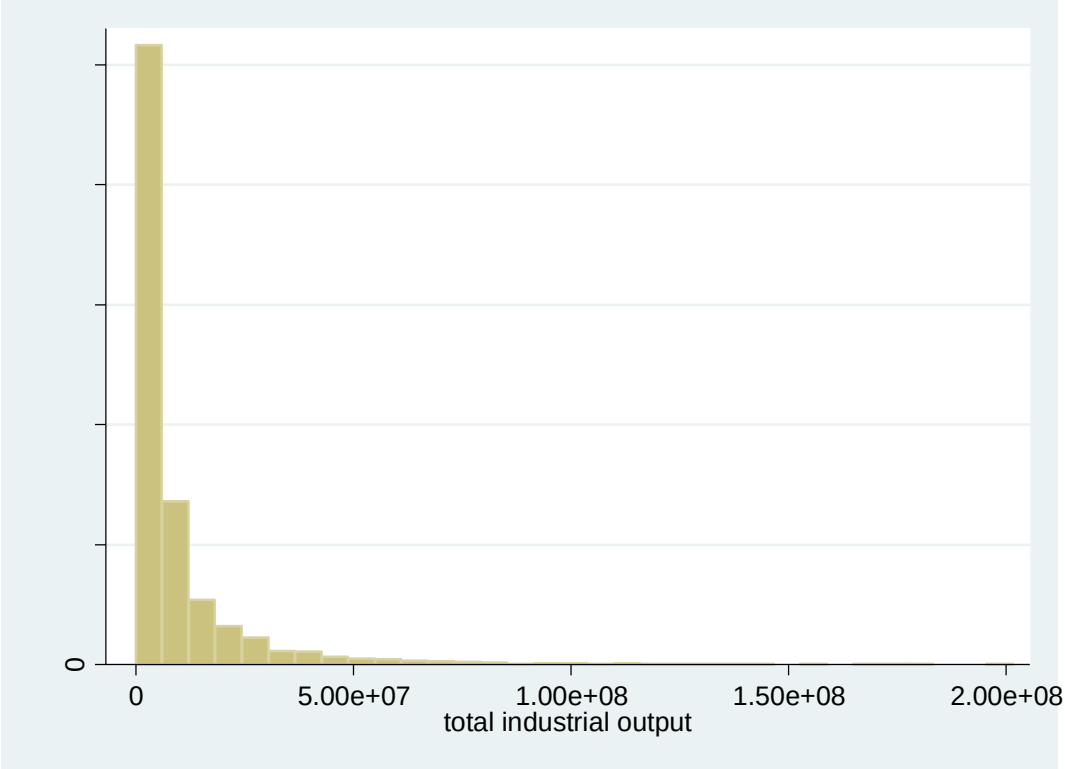
3) Share of total industrial output by foreign firms

Variable name: ind_val_fie_per
Missing observations: 123/2296
Description: Percentage (%) of industrial output by foreign (excluding HMT) firms out of total industrial output
Descriptive statistics:
Mean: 9.52
Median: 5.29
Standard deviation: 11.11
Minimum: 0
Maximum: 77.99

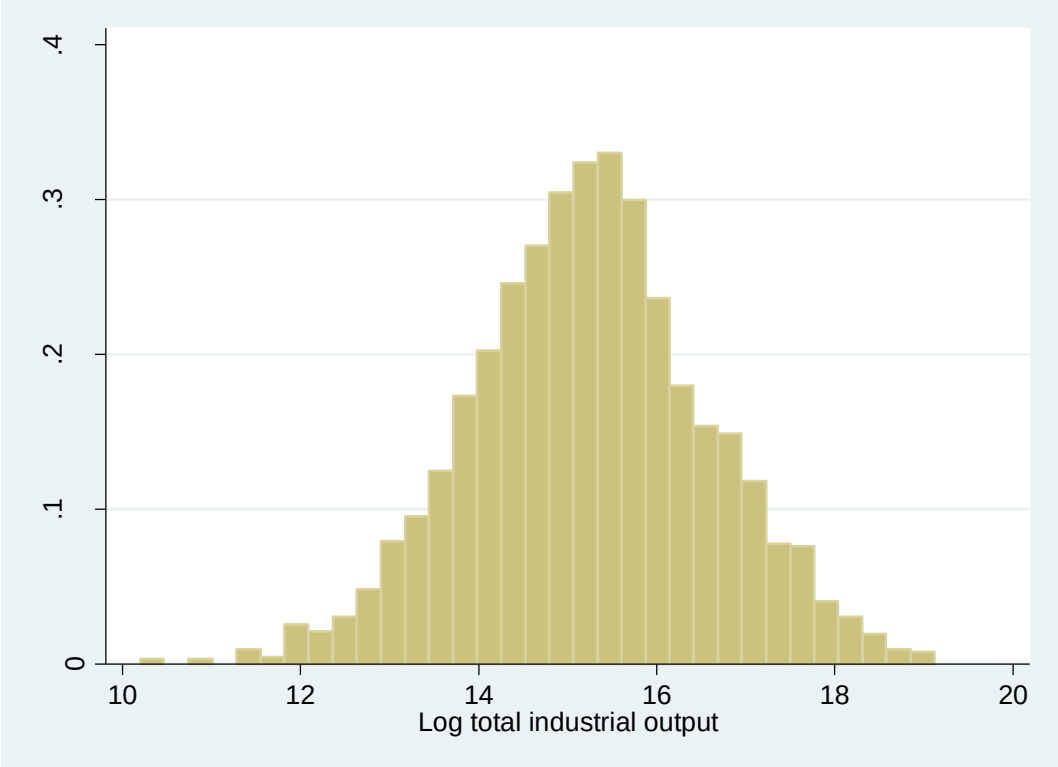


4) Total industrial output

Variable name: ind_val_real
Missing observations: 13/2296
Description: Real value (10000 yuan) of total industrial output of a city, deflated to base year 2000
Descriptive statistics:
Mean: 9827116
Median: 4201077
Standard deviation: 1.76e+07
Minimum: 26858.07
Maximum: 2.02e+08

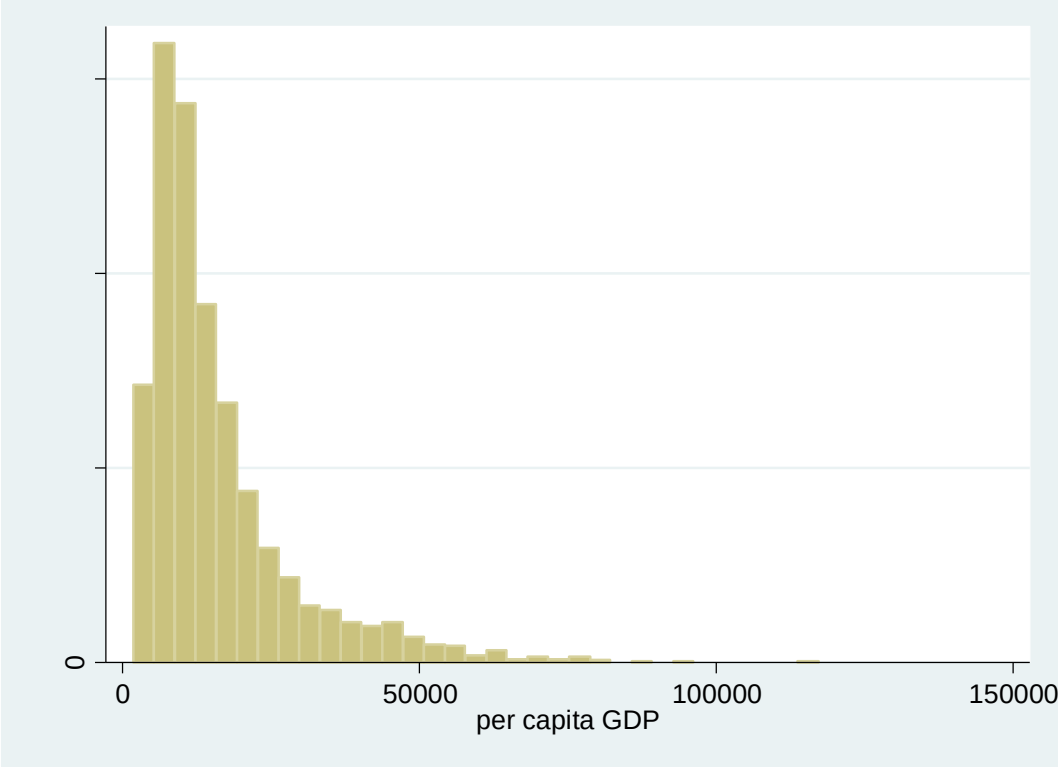


Histogram of log transformed form of the variable:

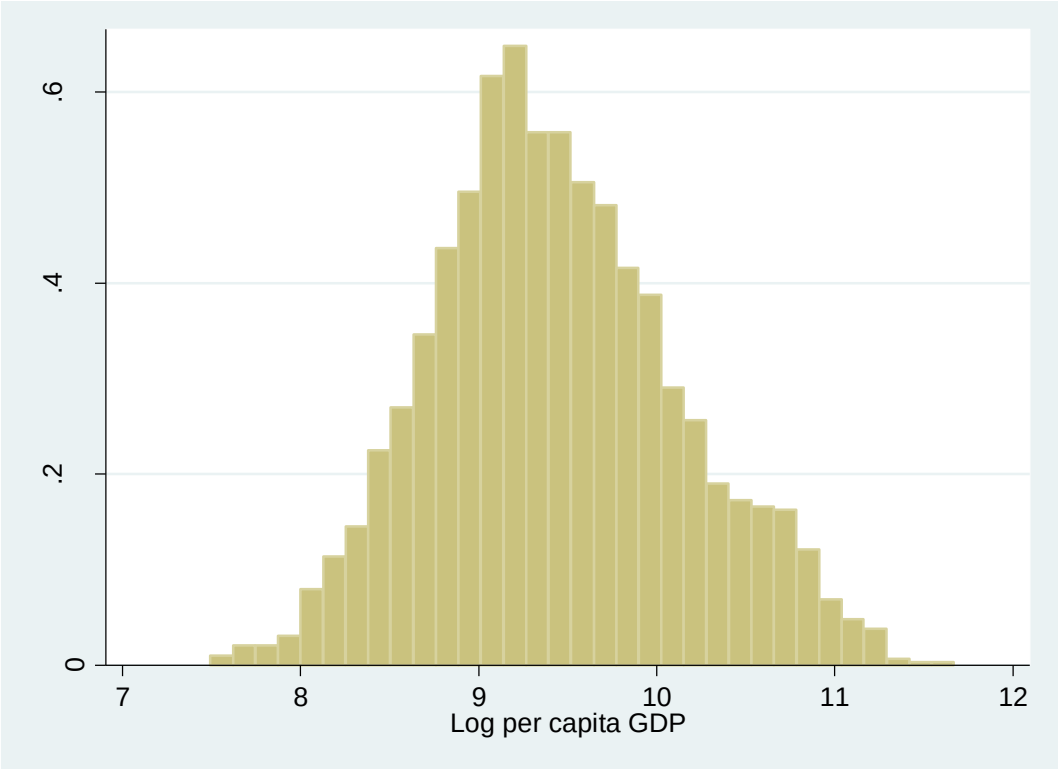


5) Real GDP per capita

Variable name: gdp_pc_real
Missing observations: 16/2296
Description: Real GDP per capita (yuan), deflated to base year 2000
Descriptive statistics:
Mean: 16067.16
Median: 12960.78
Standard deviation: 12962.27
Minimum: 1796.43
Maximum: 117250

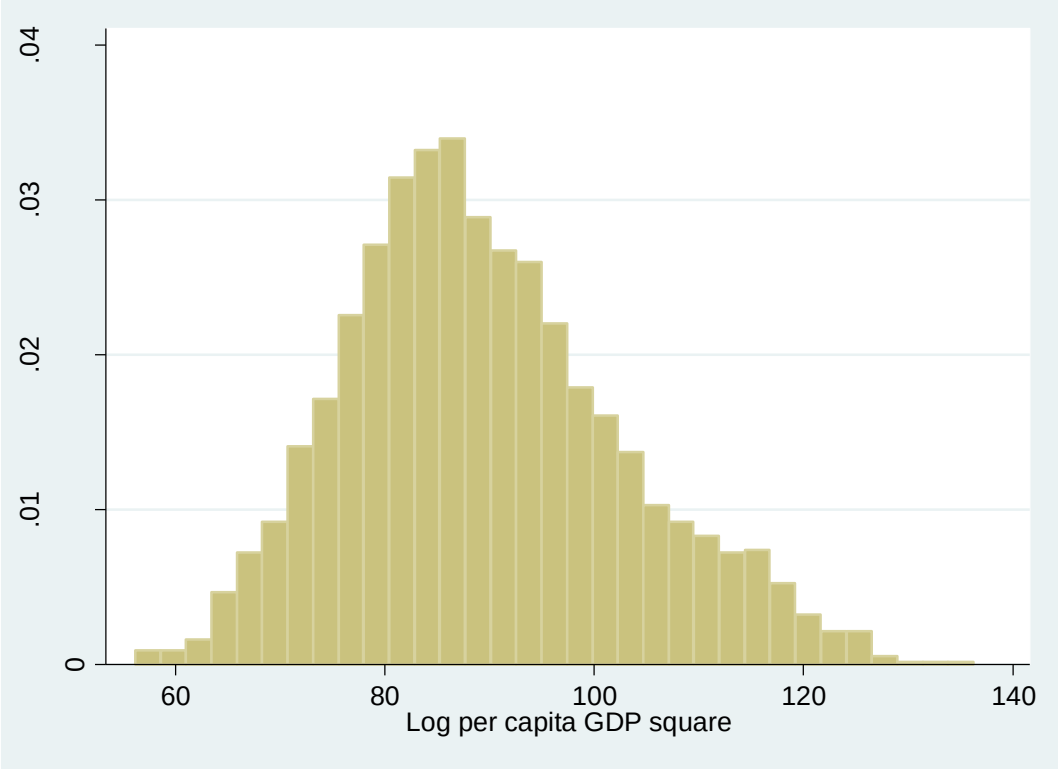


Histogram of log transformed form of the variable:



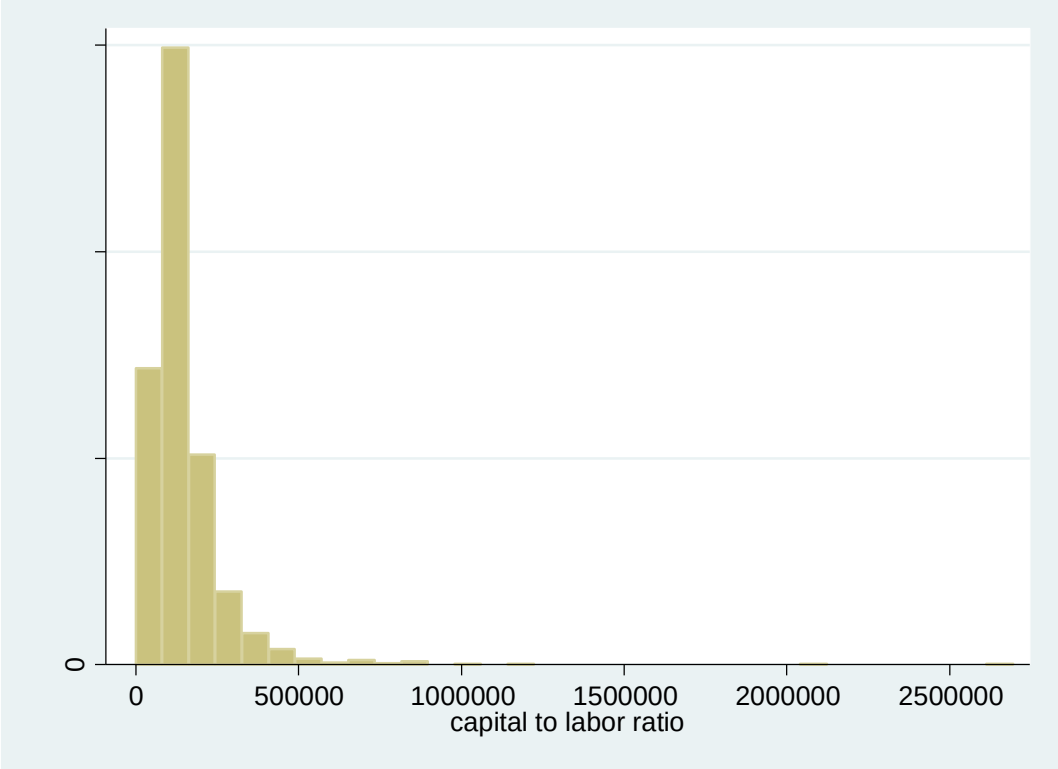
6) Real GDP per capita square

Variable name: lngdp_pc_real_sq
Missing observations: 16/2296
Description: square of the log form of real GDP per capita
Descriptive statistics:
Mean: 89.48
Median: 87.95
Standard deviation: 13.22
Minimum: 56.15
Maximum: 136.2371

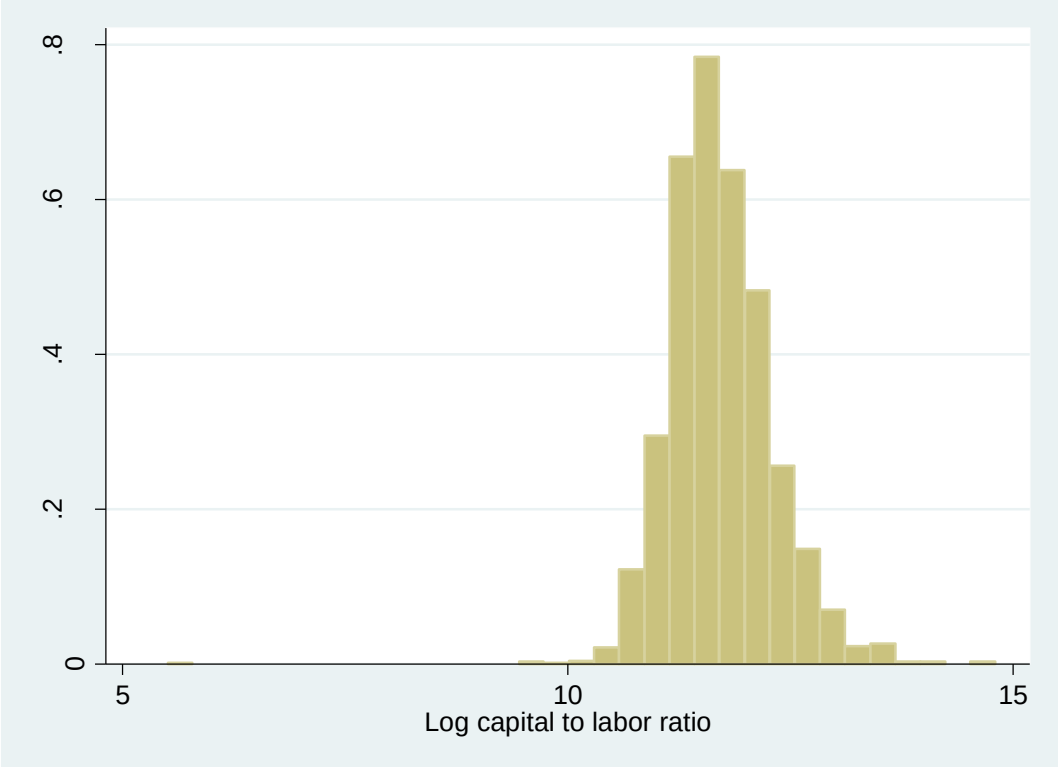


7) Capital to labor ratio

Variable name: kl_ratio_real
Missing observations: 39/2296
Description: Capital to labor ratio, measured by real net value of fixed asset (10000 yuan) per worker in industrial firms, deflated to base year 2000
Descriptive statistics:
Mean: 147933.9
Median: 116045.6
Standard deviation: 126907
Minimum: 246.18
Maximum: 2694106

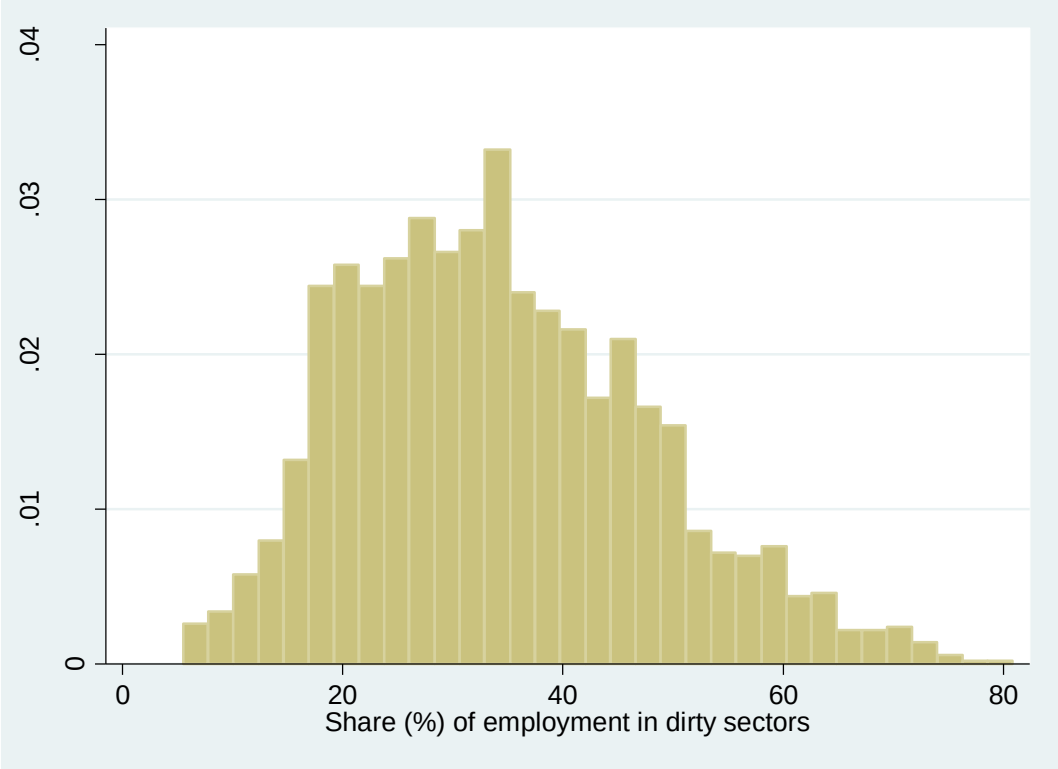


Histogram of log transformed form of the variable:



8) Share of people employed in dirty sectors

Variable name: emp_dirty_per
Missing observations: 8/2296
Description: Percentage (%) of people employed in mining, manufacturing, and utilities out of total employment level
Descriptive statistics:
Mean: 34.2
Median: 33.0
Standard deviation: 13.50
Minimum: 5.50
Maximum: 80.86

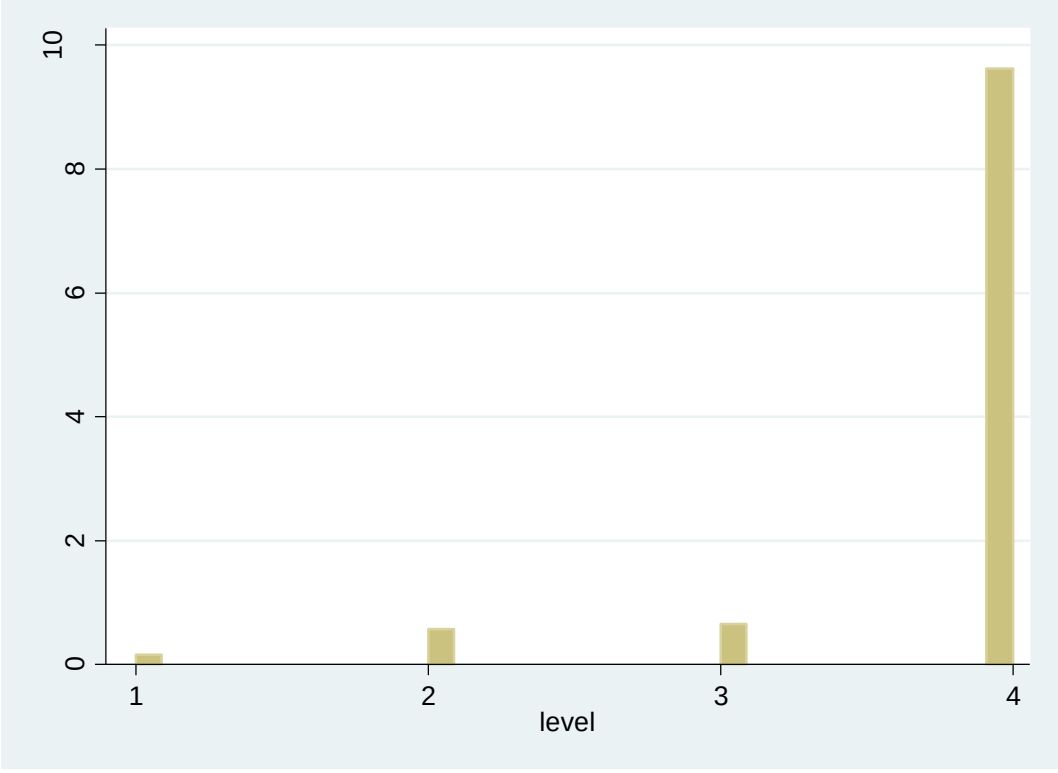


Categorical variable:

1) Administrative level

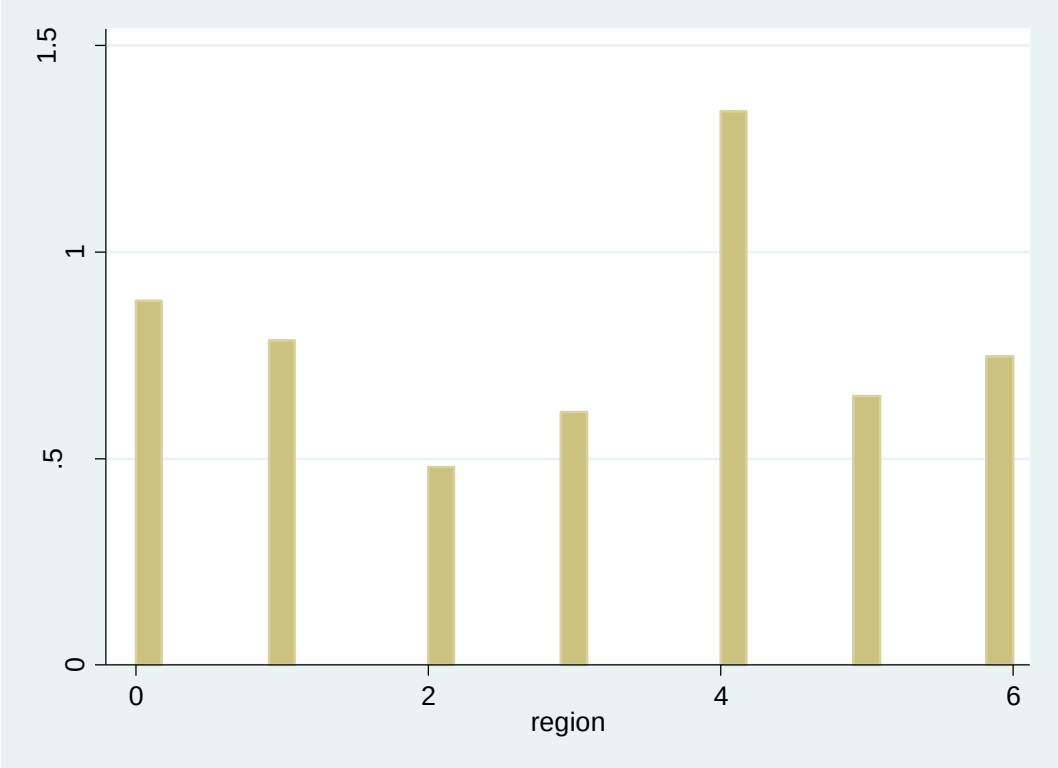
Variable name: level
Missing observations: 0/2296
Description: administrative level of a city
Coding scheme: ordered categorical
1: Municipalities
2: Sub-provincial-level cities
3: Other provincial and autonomous region capitals, all at prefecture-level
4: All other prefecture-level cities

level	Freq.	Percent	Cum.
1	32	1.39	1.39
2	120	5.23	6.62
3	136	5.92	12.54
4	2,008	87.46	100.00
Total	2,296	100.00	



2) Region
Variable name: region
Missing observations: 0/2296
Description: geographical region to which a city belongs
Coding scheme: unordered categorical
0: Southwest
1: North
2: East
3: South
4: Central
5: Northeast
6: Northwest

region	Freq.	Percent	Cum.
0	368	16.03	16.03
1	328	14.29	30.31
2	200	8.71	39.02
3	256	11.15	50.17
4	560	24.39	74.56
5	272	11.85	86.41
6	312	13.59	100.00
Total	2,296	100.00	



Tests:

Breusch-Pagan test for regression a) confirms the existence of heteroskedasticity in regression a). I therefore ran regression a) with robust standard errors to reduce this problem.

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of lnso2_emi

         chi2(1)      =  1179.26
         Prob > chi2  =   0.0000
```

Breusch-Pagan test for regression b) shows that city administrative level dummy variables and regional dummy variables have introduced heteroskedasticity to regression b). I ran a Weighted Least Square regression to reduce this problem.

```
Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: level1 level2 level3 level4 region1 region2 region3 region4
                    region5 region6 region7

         chi2(9)      =  1053.76
         Prob > chi2  =   0.0000
```