README for Data and Program

"Banking Integration and Capital Misallocation: Evidence from China"

Naide Ye
Shanghai Advanced Institute of Finance,
Shanghai Jiao Tong University
ndye.20@saif.sjtu.edu.cn

Dongmin Kong
School of Economics,
Huazhong University of Science and Technology
kongdm@mail.hust.edu.cn

1. The Structure of the Data and Program

The folder includes several key components to which we refer below.

--Program.do

--A series of dta files for empirical analyses

2. Details about Datasets

Our data consists of aggregate-level data and firm-level data. We make our aggregate-level data public. Given the proprietary nature of the firm-level data, we are unable to make the data public.

2.1 Dataset#1

Dataset name: *prov_panel*.dta

This dataset is used to generate Table 2. The dataset is a province-year panel dataset including growth rate of GDP (*gdpgrowth*), log of deposits per capita (*logdeppc*), log of investments per capita (*loginvpc*), log of value added from industrial firms per capita (*logindusvapc*), log of value added from financial industry per capita (*logfinvapc*), log of GDP per capita (*loggdppc*), standard deviation of logarithm of firm MRPK (*logmrpksd*). We build this dataset mainly based on *China Statistical Yearbook*.

2.2 Dataset#2

Dataset name: *city_panel*.dta

This dataset is used to generate Table 3, Table 4, Figure 4 and Figure 5. The dataset is a city-year panel dataset, which includes key variables as follows: log investments (*loginv*), log of investments per capita (*loginvpercap*), log of deposits per capita (*logsavingpercap*), log of GDP per capita (*loggdp_per_capita*), log of population (*logpop*), the share of High MRPK firms (*highmrpk_share*). We extract data mainly from the City Statistical Yearbook to construct these variables, and construct variable *highmrpk_share* based on the Annual Survey of Industrial Firms (ASIF) and the Annual Tax Survey (ATS).

2.3 Dataset#3

Dataset name: *firm_panel*.dta

This dataset is used to generate Table A1, Table 5, Figure 6, Table 6, Table 7, Table 9, Table 10, Table 11, Table 12, Table 13, Table 14 and Table 15. The dataset is a firm-year panel dataset including firm's age (*age*), log of firm's fixed assets (*logcapital*), log of firm's MRPK (*logmrpk*), log of long-term debt (*logdebt*), log of capital expenditure (*loginv*), indicator for non-SOE (nsoe), indicator for higher MRPK prior to 2004 (*highmrpkpre04*), capital size quartile bins (*capquartile*), two-digit industry (*ind2digit*), four-digit industry (*ind4digit*), indicator for newly added loans (*lending_extensive*). We construct these variables based on the Annual Survey of Industrial Firms (ASIF) and the Annual Tax Survey (ATS).

2.4 Dataset#4

Dataset name: *control_vs_peer*.dta

This dataset is used to generate Table C1. We construct this dataset based on Dataset *firm_panel*.dta but only keep the "peer" and "control" group in the sample.

2.5 Dataset#5

Dataset name: *city_matching*.dta

This dataset is used to generate Figure C2 to show the effectiveness of PSM. This dataset is a city-level cross-sectional data. Key variables include log of GDP per capita (*loggdppc*), log of loan per capita (*logloanpc*), log of deposit per capita (*logsavingpc*), fiscal budget scaled by local GDP (*fiscalbudget*), log of investments per capita (*loginvpc*), mean intensity of night light (*lightintensity*), share of output of the second industry (*industrialshare*), share of output of the third industry (*tertiaryshare*), growth rate of local GDP (*gdpgrowth*) and amount of foreign direct investment scaled by local GDP (*fdi*). We construct this dataset mainly based on City Statistical Yearbook.

2.6 Dataset#6 and #7

Dataset name: *firm_clean_without_matching*.dta and *firm_clean_matching*.dta

During the post-integration period, some of these integrated banks may expand branches into "additional" prefectures. Firms located in "additional" prefectures might experience shocks from "original" prefectures through new bank branches. To eliminate the impact of these shocks, we exclude the "additional" prefectures from our sample. Thus, we construct the data file *firm_clean_without_matching*.dta. Based on this dataset, we perform PSM and retain only matching cities and then construct the dataset *firm_clean_matching*.dta.

2.7 Dataset#8 and #9

Dataset name: *prov_ind_dispersion*.dta and *city_ind_dispersion*.dta

These datasets are used to generate Table 15. These datasets are province(city)-industry-year panel data. Key variables include log MRPK at 80% percentile minus that at 20% percentile (*disper_80_20*), log MRPK at 75% percentile minus that at 25% percentile (*disper_75_25*), and variance of log MRPK (*logmrpk_var*).

2.8 Dataset#10

Dataset name: *local_bank_hhi*.dta

This dataset is used to generate Figure B2 and Figure 7. This dataset is a province-year panel dataset. Key variables in this dataset include local banking geographical concentration index (*weighted_hhi*), standard deviation of firm MRPK (*logmrpk_scale_prov_sd*) and log of GDP per capita (*loggpc_real*).