ECONOMETRICS 1 - HOMEWORK

PILI Pierre

December 7, 2023

1 The Basics

1.1 Create a table of descriptive statistics of the variables in the dataset.

var	median	mean	min	max	sd	NAs
agro_emp	18.6	25.1	0.1	86.3	22.3	30
bribery	11.7	17.0	0.0	67.1	14.7	87
gfce	16.5	17.7	5.1	62.9	8.4	36
literacy	93.0	83.6	24.2	100.0	19.3	61
log_gdp	9.4	9.3	6.6	11.6	1.2	22
pop_total	6.2e+06	3.4e+07	1.1e+04	1.4e+09	1.4e+08	2
self_emp	35.0	40.9	0.4	94.8	27.0	30
stocks	6.4	28.8	0.0	538.7	66.8	131
sample_size	715.1	3.6e+03	120.1	1.4e+05	1.3e+04	2

Table 1: Descriptive statistics

The data is made of 217 countries and certain variables contains a significant amount of missing values (NAs) (See 1). Bribery is the worst variable with up to 87 NAs. Using this variable in a regression would yield a low precision as the size of the sample is quite small. This table demonstrates how different countries are, indeed the min-max interval is rather close to one for rate variables and the standart deviations (sd) are very significant compared to the mean value of each variable.

1.2.a Is it the case that self-employment is correlated with how rich a country is (in-terms of log GDP per-capita)?

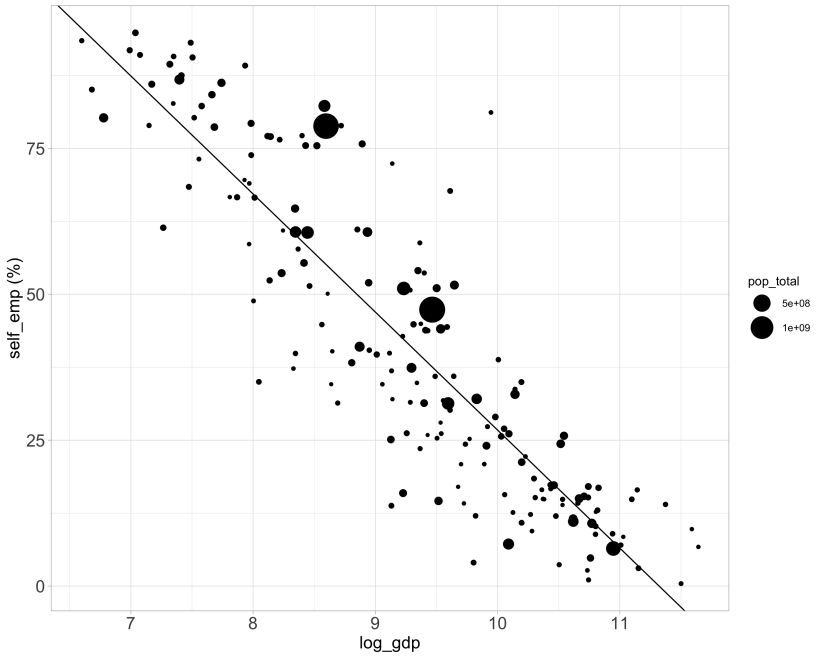

Figure 1: Self employment rate with respect to GDP

Figure 1 shows a clear negative relationship between the share of self employment and gdp. The empirical correlation coefficient is equal to -0.89 which is very close to -1. The anticorrelation is very strong.

1.2.b Is it the case that countries with higher share of employment in agriculture also have higher self-employment rates?

As in the previous question, Figure 2 shows a clear positive relationship between the share of self employment and share of employment in the agricultural sector. The empirical correlation coefficient is equal to 0.91 which is very close to 1. The correlation is very strong.

1.2.c Present a bar graph comparing the mean self-employment rates in each of these 3 literacy-based categories of countries.

Figure 3 seems to demonstrate a negative relationship between the literacy category of the population and the self employment rate.

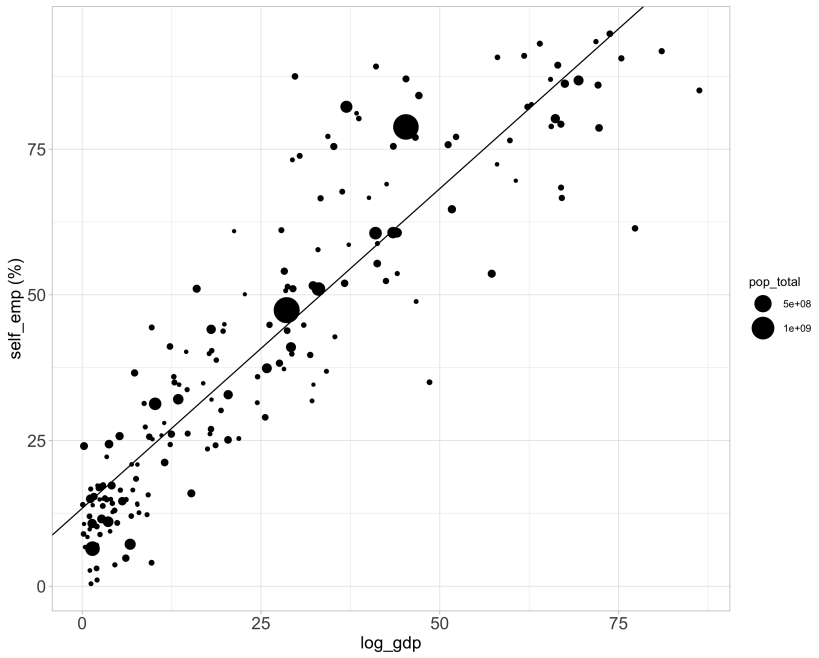

Figure 2: Self employment rate with respect to employment share in the agricultural sector

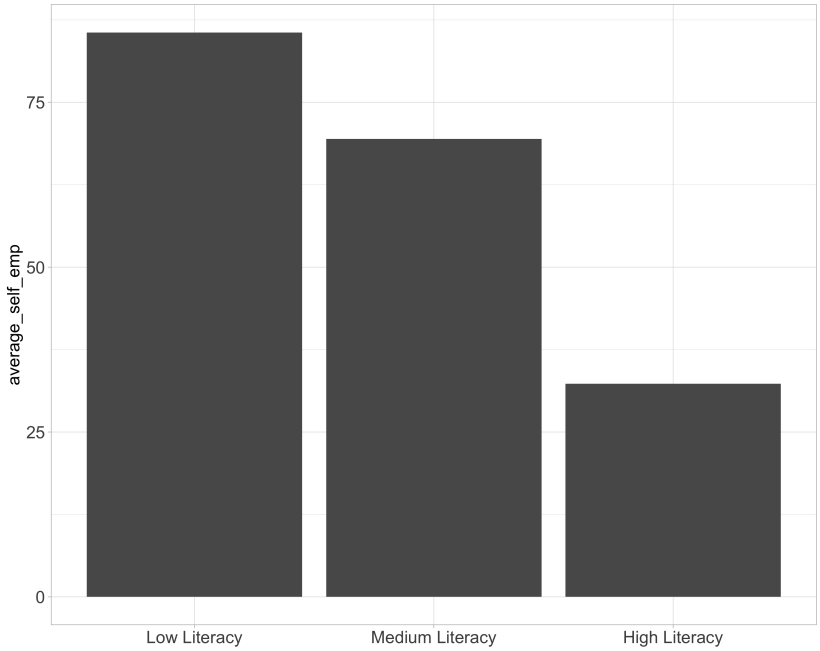

Figure 3: Average self employement rate as a function of the literacy category

1.3 Estimate the model parameters described using OLS, report your results and summarize them

Denoting this linear model as (1) and estimating it yields the results displayed in Table 2 (first column). This OLS estimation was based on only 143 observations, which is far from 217. This raises a question, are NAs equally distributed across the sample ? That being aknowledged, this regression seems to be very significant. The overall significance is very high as the F-stat demonstrates. Every of the three coefficients are significantly different from zero at the 1% level. The sign of the variables is in line with the previous discussion in question 1.1. According to this estimation, a 1% increase in GDP corresponds to a 6.5% decrease in the share of self employement. A 1% increase in the literacy rate is associated with a 0.3% decrease in the dependent variable, while a 0.59% increase in self employment shares can be explained by a 1% increase in the share of the agricultural sector.

Table 2: Linear Regressions - Exercise 1

	Dependent variable:	
	self_emp	
	(1)	(2)
log_gdp	-6.506^{***}	-5.520
	(1.755)	(4.042)
literacy	-0.313^{***}	-0.358^{**}
	(0.070)	(0.175)
agro_emp	0.592^{***}	0.628^{***}
	(0.080)	(0.176)
gfce		-0.922^{**}
		(0.380)
stocks		0.110^{*}
		(0.061)
bribery		-0.111
		(0.156)
Constant	113.219^{***}	121.953^{***}
	(16.361)	(41.265)
Observations	143	49
R^2	0.845	0.828
Adjusted R^2	0.841	0.804
Residual Std. Error	10.574 (df = 139)	9.262 (df = 42)
F Statistic	252.152^{***} (df = 3; 139)	33.720^{***} (df = 6; 42)
Note:		*p<0.1; **p<0.05; ***p<0.01

1.4 Describe how you could estimate β_3 using a 3-step procedure based on the Frisch-Waugh Theorem.

The Frisch-Waugh theorem can be used for estimating specific coefficients in a multiple regression model while controlling for other variables. Here are the simplified three steps:

Step 1:
Identify control variables, here log_gdp and literacy and regress it on the dependent variable to get the residuals $\hat{r}$.

$$\text{self\_emp} = \alpha_0 + \alpha_1 log\_gdp + \alpha_2 literacy + r$$

Step 2:
Regress the variable of interest on the control variables and get the residuals $\hat{u}$.

$$\text{agro\_emp} = \gamma_0 + \gamma_1 log\_gdp + \gamma_2 literacy + u$$

Step 3:
Regress $\hat{r}$ on $\hat{u}$ and get the coefficient of interest β_3.

$$\hat{r} = \beta_0 + \beta_3 \hat{u} + \epsilon$$

The coefficient β_3 is the estimate of the parameter for agro_emp after controlling for log_gdp and literacy.

1.5 Implement this 3-step procedure and compare your estimate and standard error of β_3

2 Heteroskedasticity & Monte Carlo Simulations

3 Instrumental Variables

3.1

Table 3 shows the descriptive statistics of the table made of 900 hundred municipalities. First, the data is remarkably clean as their is no NA values in it.

var	median	mean	min	max	sd	NAs
business_crea	0.1	0.1	5.6e-02	0.3	3.1e-02	0
nb_crimes	73.0	149.8	9.0	1.0e+04	451.3	0
nb_households	9.8e+03	1.9e+04	3.7e+03	1.4e+06	5.3e+04	0
pop	2.0e+04	3.5e+04	1.0e+04	2.2e+06	8.8e+04	0
income	2.0	2.9	1.0	6.0	1.1	0
crime_rate	7.2e-03	7.9e-03	1.0e-03	3.0e-02	4.1e-03	0

Table 3: Descriptive statistics

Table 4: Results

	Dependent variable:
	crime_rate
business_crea	0.018***
	(0.005)
log(pop)	0.001***
	(0.0002)
income	0.00001
	(0.0001)
com_typeC	−0.003***
	(0.0003)
com_typeI	−0.002***
	(0.001)
Constant	0.001
	(0.002)
Observations	899
R^2	0.150
Adjusted R^2	0.145
Residual Std. Error	0.004 (df = 893)
F Statistic	31.568*** (df = 5; 893)
Note:	*p<0.1; **p<0.05; ***p<0.01