# ECONOMETRICS 1 - HOMEWORK

PILI Pierre

January 19, 2024

## 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)?

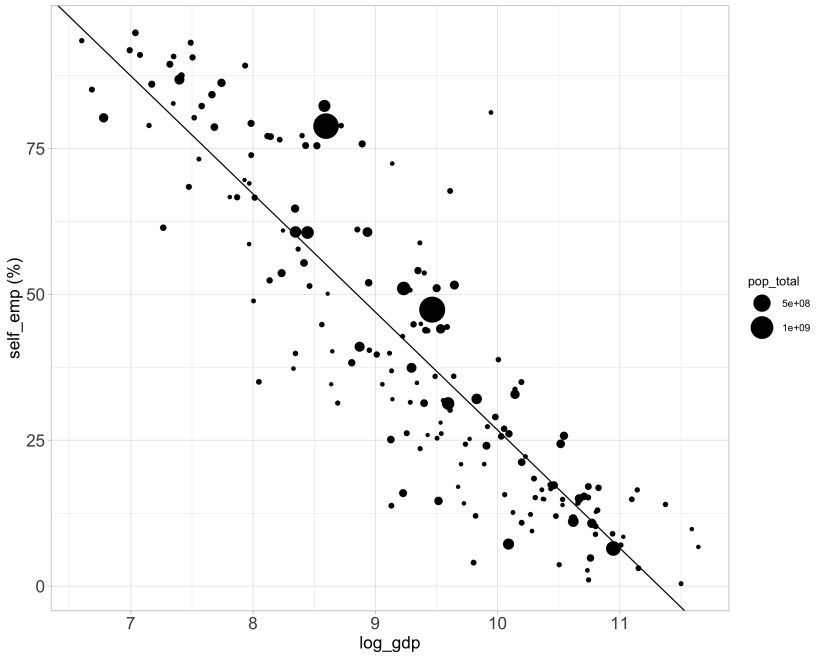


Figure 1: Self employment rate with respect to GDP

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.

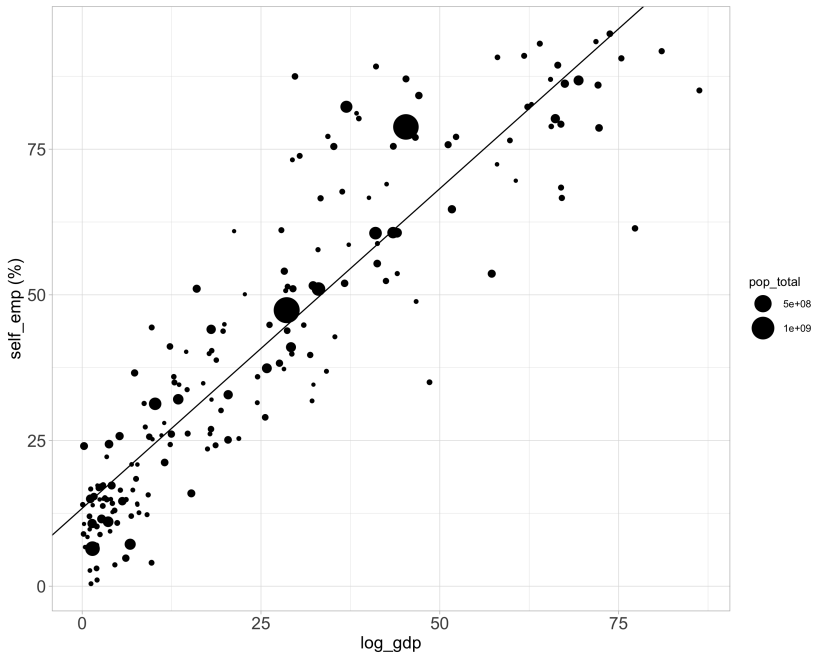


Figure 2: Self employment rate with respect to employment share in the agricultural sector

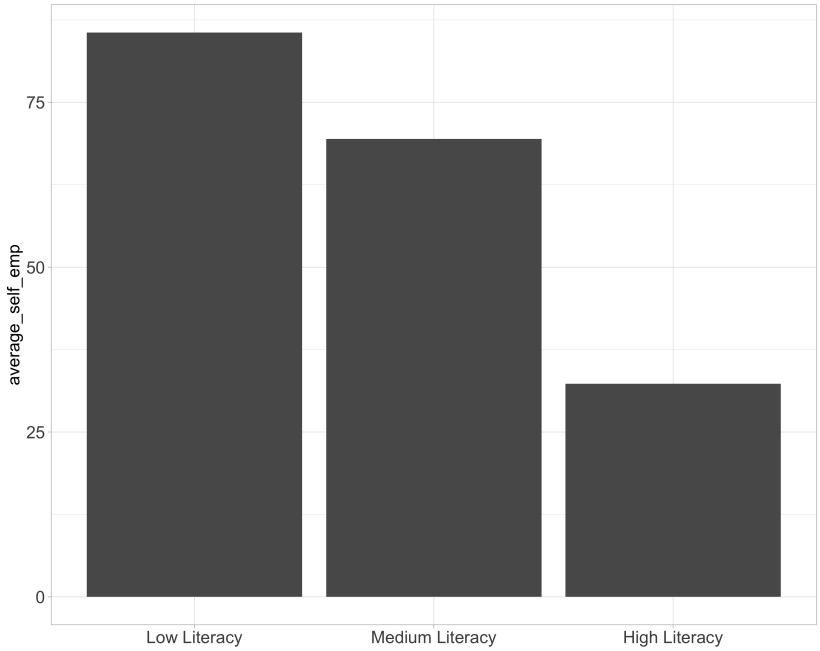


Figure 3: Average self employement rate as a function of the literacy category

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

Denoting this linear model as (1) (we shall refer to it as the "simple estimation" in the rest of this section) 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 signs of the variables are 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: Simple and Extended Models - 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 $\beta_3$ 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 $\hat{r}$. The residuals of this regression is the part of self_emp that is not explained by the log_gdp and literacy variables.

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

**Step 2:**
Regress the variable of interest on the control variables and get the residuals $\hat{u}$. The residuals of this regression is the part of agro_emp that is not explained by the log_gdp and literacy variables.

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

**Step 3:**
Regress $\hat{r}$ on $\hat{u}$ and get the coefficient of interest $\beta_3$.

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

The coefficient $\beta_3$ 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 $\beta_3$.

Table 3 and Table 4 display the results of the 3-step procedure previously described, where $\hat{r}$, $\hat{u}$ are respectively denoted as r_hat and u_hat. The first regression yields coefficients that are very different from the first estimation (first column of Table 2). This is an expected result, indeed as agro_emp is assumed to be part of the data generating process, the first step estimation is inconsistent because of endogeneity issues. The higher the correlation between agro_emp and control variable, the larger is the discrepency between coefficients from the first step of the Frisch-Waugh procedure and the simple estimation. The third regression yields an estimation for $\beta_3$ and its standard error which are very close to what we got in the simple estimation. Those results were expected and a direct implication of the Frisch-Waugh theorem.

Table 3: First and Second Regressions in 3-step Procedure - Exercise 1

| | *Dependent variable:* | |
|---|---|---|
| | self_emp | agro_emp |
| | (1) | (2) |
| log_gdp | −15.784*** | −15.670*** |
| | (1.453) | (1.308) |
| literacy | −0.377*** | −0.108 |
| | (0.081) | (0.073) |
| Constant | 219.877*** | 180.127*** |
| | (9.247) | (8.326) |
| Observations | 143 | 143 |
| $R^2$ | 0.783 | 0.743 |
| Adjusted $R^2$ | 0.780 | 0.739 |
| Residual Std. Error (df = 140) | 12.454 | 11.215 |
| F Statistic (df = 2; 140) | 252.747*** | 201.952*** |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 | |

Table 4: Third Regression in 3-step Procedure - Exercise 1

| | *Dependent variable:* |
|---|---|
| | r_hat |
| u_hat | 0.592*** |
| | (0.079) |
| Constant | 0.000 |
| | (0.878) |
| Observations | 143 |
| $R^2$ | 0.284 |
| Adjusted $R^2$ | 0.279 |
| Residual Std. Error | 10.499 (df = 141) |
| F Statistic | 56.008*** (df = 1; 141) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

#### 1.6.a Estimate a linear model that includes the expanded set of covariates using OLS. Report your results and interpret them

Denoting this model as (2) (we shall refer to it as the "extended model"), the results are displayed in Table 2 along with the simple model estimation. The common coefficients among both model are in line with each other regarding values and errors. The bribery variable is not significantly different from zero at the 10%. The stocks variable is significantly different from zero at the 10% level which is not a very strong result. The gfce variable is significantly different from zero at the 5% level and means that a 1% increase in the value of stocks traded as a share of GDP is associated with a roughly 1% increase in the share of self employement. In other words, the larger the financial sector, the larger the share of self employement in the economy.

#### 1.6.b How many observations was this model estimated on? Is it identical or different from the number of observations in question 3 and why?

The first thing to notice by comparing the simple and extended models is that the sample size is much smaller for the extended model as the individuals with at least one NA among the 6 variables were removed. The

immediate consequence is that the simple model is roughly twice as precise as the extended one.

**1.6.c We are told that many of these variables, including the variable on self-employment rates, are measured using sample surveys in each country. This implies that the variance of each observation is directly proportional to the sample size used in each country. What does this imply for the OLS estimator?**

The statement that the variance of each observation is directly proportional to the sample size implies heteroscedasticity in the data. Heteroscedasticity violates one of the assumptions of OLS regression, which assumes that the variance of the error term is constant across all levels of the independent variables. Indeed, testing for heteroscedasticity in the simple model using a White test leads to reject the homoskedascity assumption at the 10% level with a p-value of 0.058. Running the White test on the extended model yields no proof of heteroscedasticity mostly because the sample is not large enough. Assuming that the variance of each observation is directly proportional to the sample size, it is enough to intensify the variables and divide them all by the sample size. What we obtain is thus a sphericalized model which should not bear any heteroscedasticity issues. Table 5 displays the results of the sphericalized estimation. The coefficient $\beta_3$ is significantly different from zero at the 1% level and positive but its value is very different from the extended OLS. This was an expected result, in case of heteroscedasticity, the OLS estimated variance has no reason to be true as the variance-covariance matrix of the noise used to compute it is wrongly specified.

**1.7 What can you conclude about the factors driving differences in self-employment rates at the macro-level based on your above answers? Can these results be interpreted as causal?**

It appear from the previous results that the main effect on self-employment comes from the size of the economy and then from the specialization of this economy (size of the agricultural sector). This was expected as a bigger economy comes with organization changes and larger firms thus decreasing the self employment rate. It seems very intuitive and the correlation is indeed very strong. However one should be very careful while underlying causal effects and inspect the issue further by running new experiments.

Table 5: Weighted Least Squares Extended Estimation - Exercise 1

| | *Dependent variable:* |
|---|---|
| | self_emp |
| log_gdp | −0.585 |
| | (1.698) |
| literacy | 0.239 |
| | (0.186) |
| agro_emp | 1.043*** |
| | (0.107) |
| gfce | −0.441** |
| | (0.193) |
| stocks | 0.383 |
| | (0.288) |
| bribery | 0.404*** |
| | (0.130) |
| Constant | −0.001 |
| | (0.002) |
| Observations | 49 |
| $R^2$ | 0.929 |
| Adjusted $R^2$ | 0.918 |
| Residual Std. Error | 0.007 (df = 42) |
| F Statistic | 91.073*** (df = 6; 42) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

# 3 Instrumental Variables

### 3.1 Compute the crime rate per household. Generate a table of descriptive statistics of all the variables that have been introduced so far. Comment on the descriptive statistics.

| var | median | mean | min | max | sd | rsd |
|---|---|---|---|---|---|---|
| business_crea | 0.1 | 0.1 | 5.6e-02 | 0.3 | 3.1e-02 | 0.2 |
| nb_crimes | 73.0 | 149.8 | 9.0 | 1.0e+04 | 451.3 | 3.0 |
| nb_households | 9.8e+03 | 1.9e+04 | 3.7e+03 | 1.4e+06 | 5.3e+04 | 2.9 |
| pop | 2.0e+04 | 3.5e+04 | 1.0e+04 | 2.2e+06 | 8.8e+04 | 2.5 |
| income | 2.0 | 2.9 | 1.0 | 6.0 | 1.1 | 0.4 |
| crime_rate | 7.2e-03 | 7.9e-03 | 1.0e-03 | 3.0e-02 | 4.1e-03 | 0.5 |

Table 6: Descriptive statistics

Table 6 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. For a clearer view I displayed the relative standard deviation (rsd) defined as the standard deviation divided by the mean.

### 3.2 Estimate the model in equation (7) using OLS.

The coefficient of interest that results from the OLS (see Table 7) is significantly different from zero at the 1% level, and positive. This is very unexpected as it would mean that the better the shape of the economy, the higher the crime rate. This result could derive from endogeneity issues as an ommited variable bias for instance. The fact that the coefficient of interest is positive while we expect it to be negative means that the endogeneity issues are more than just measurment error as this kind of problem leads underestimations (in absolute value) of the true coefficient.

### 3.3 After estimating the model in equation (7), you wonder whether enodegeneity is a problem. Why could the variable of interest, the business creation growth rate, be endogenous?

As mentioned in the previous question, the fact that the estimation for the coefficient of interest is positive leads to thinking about an ommited variable bias. For instance, both the shape of the economy and the crime rate could be related to the level of inequality in the area. To inspect this hypothesis,

Table 7: Ordinary Least Square Estimation - Exercise 3

| | *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 |

we used the iqr_income variable which is defined as the difference between the first and third quartiles of income. This is a proxy for the local level of inequality. Figure 4 displays the crime rate and the business creation growth rate against the iqr_income variable. Even though the result is not cristal clear, we still find some positive correlation for both the dependant and the explanatory variable. This is a source of enodegeneity. The coefficient of correlation with iqr_income are respectively 0.23 and 0.082 for the business creation rate and the crime rate.

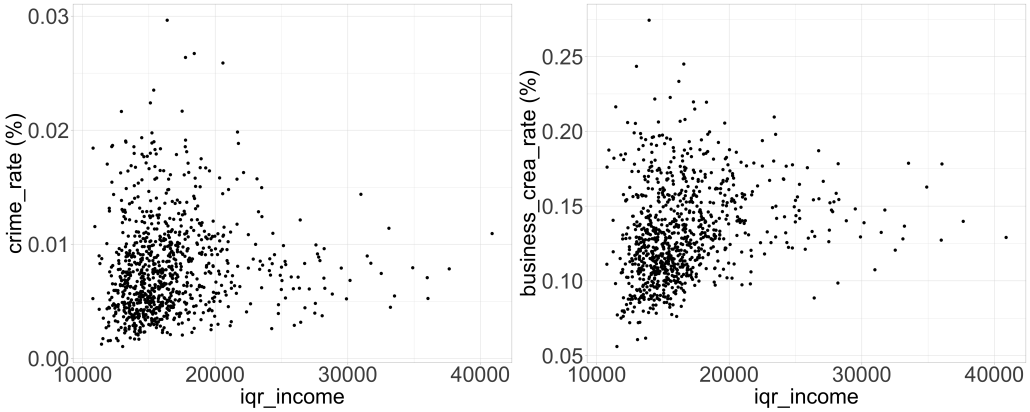


Figure 4: Crime rate and business creation growth rate with respect to inequality indicator

#### 3.4.a Discuss the validity of all potential instruments with respect to the relevance condition.

For an instrumental variable to be relevant, they should have no direct impact on the dependant variable (orthogonality condition). Are there any reason to think that one of the three variables impact local crime rate ? It seems very unlikely for the motorway variable, unless a majority of crimes happen to occur during motoraces, which seems ridiculous. If we assume that the crime rate is driven by the living condition of a population, it seems unlikely that changing the political party of the mayor *ceteris paribus* would have any impact on the crime rate. As for the tax rate variable, I would like to be more cautious. If, as I believe it is, the inequality level is part of the data generating process, then it could be that the local tax rate has an impact on the local level of (perceived) inequality and would impact the crime rate through this ommited variable. One should inspect the literature on the subject to build intuition about such a matter.

### 3.4.b Use the correlation between the business ce rate and the three variables to assess their potential strength as instruments. Does one seem better suited as an instrument than the other? Why?

After computing the correlation between the instruments and the business creation growth rate, it appears that the motorway variable seems to be better suited than the other with correlation of 0.295 compared to 0.165 for the tax_rate and -0.005 for the political party (see Figure 5). The tax_rate could still be used as an instrument with a correlation of 0.165 while the right variable would not add any information to the model.

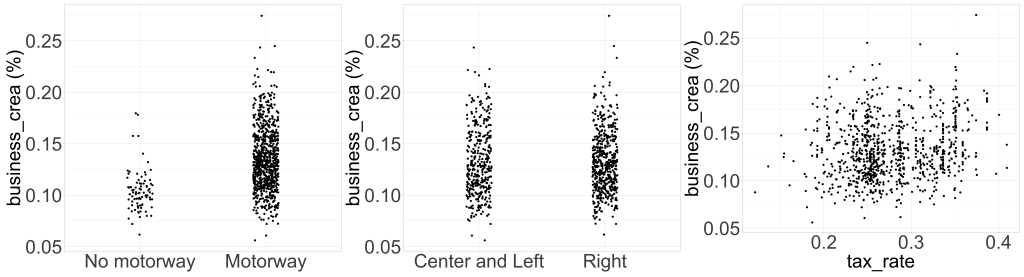


Figure 5: Business creation growth rate against instruments

### 3.4.c Thoroughly discuss the validity of the IV candidates regarding the exclusion restriction.

We have discussed the exogeneity condition in the question 3.4.a which can be summarized as such : there are no reasons to think the the party and motorway variables are part of the unexplained variations of the dependent variable while the tax rate variable could impact the level of inequality and thus impact the crime rate. We cannot check the orthogonality condition as it relies on the distribution of the noise which we cannot access. To inspect the exclusion restriction, we can run a regression adding the instrumental variables to the equation of interest and check that the coefficients are not significantly different from zero (see Table 8). It appears that the only instrumental variable that does not explain any variation in the dependant variable is the party variable. It is not very convenient as the party variable is also the weakest instrumental variable according to 3.4.b. Although I have no explanation for the impact of the motorway variable, the coefficient before the tax_rate does not seem to confirm our analysis about the impact of this variable through the level of inequality as it is positive. Maybe the tax_rate is perceived as unfair and actually increases the tension between

the State and the population.

Table 8: Extended model - Exercise 3

| | *Dependent variable:* |
|---|---|
| | crime_rate |
| business_crea | 0.012*** |
| | (0.005) |
| log(pop) | 0.0003* |
| | (0.0002) |
| income | 0.0004*** |
| | (0.0001) |
| com_typeC | −0.002*** |
| | (0.0003) |
| com_typeI | −0.001** |
| | (0.001) |
| motorway | 0.001*** |
| | (0.0004) |
| partyRight | −0.001 |
| | (0.001) |
| tax_rate | 0.029*** |
| | (0.003) |
| Constant | −0.006*** |
| | (0.002) |
| Observations | 899 |
| $R^2$ | 0.261 |
| Adjusted $R^2$ | 0.254 |
| Residual Std. Error | 0.004 (df = 889) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |

**3.5.a Run 3 first-stage regressions, each using one of the above instruments separately. Report the 3 sets of results in the same table. Which instrument seems to be the strongest?**

The results are displayed in Table 9 and correspond to what we displayed in Figure 5. The motorway variable seems to be the strongest instrument with a coefficient significantly different from zero at the 1% level whereas the tax_rate is only significant at the 5% level and the party variable is not significant whatsoever.

**3.5.b Extract the fitted values of any one of these first-stage regressions and use them to run a second-stage regression. Assuming that the instrument is exogenous, are your results reliable? Why?**

The results are displayed in Table 10 along with the ivreg 2SLS regression. Although the coefficient of interest is significant at the 1%, we have seen in Table 8 that the motorway variable was part of the data generating process, thus even assuming the exogeneity condition is not enough to rely on those results.

**3.5.c For the instrument chosen in part (b), instead of manually running the 2 stages, implement the IV estimator using the ivreg package. Is your answer the same or different to the answer in part (b) and why?**

The coefficients and variances are the almost identical as in the previous question because the computation follows the same process (see Table 10). That aside, the $R^2$ coefficient is negative which is a sign that the model is either wrong or overfitted.

**3.5.d Use all instruments you think are valid to implement the IV estimator using the ivreg package and report your results. How different are the results from part (c)?**

The two estimators that are correlated enough with the endogenous variable are the motorway and tax_rate variables (see Table 9), assuming the orthogonality condition for both of them, they are the only estimators to use in a 2SLS regression. The results are displayed in Table 11 and we get an even higher coefficient for the growth rate of the economy which would

Table 9: First Stages - Exercise 3

| | *Dependent variable:* | | |
|---|---|---|---|
| | business_crea | | |
| | motorway | party | tax_rate |
| log(pop) | 0.011*** | 0.011*** | 0.011*** |
| | (0.001) | (0.001) | (0.001) |
| income | −0.001 | −0.001 | −0.0001 |
| | (0.001) | (0.001) | (0.001) |
| com_typeC | −0.032*** | −0.034*** | −0.033*** |
| | (0.002) | (0.002) | (0.002) |
| com_typeI | −0.030*** | −0.033*** | −0.032*** |
| | (0.004) | (0.004) | (0.004) |
| motorway | 0.014*** | | |
| | (0.003) | | |
| partyLeft | | −0.0003 | |
| | | (0.004) | |
| partyRight | | 0.001 | |
| | | (0.004) | |
| tax_rate | | | 0.039** |
| | | | (0.019) |
| Constant | 0.026** | 0.034*** | 0.024* |
| | (0.012) | (0.013) | (0.013) |
| Observations | 899 | 899 | 899 |
| $R^2$ | 0.338 | 0.322 | 0.325 |
| Adjusted $R^2$ | 0.334 | 0.317 | 0.321 |
| Residual Std. Error | 0.025 (df = 893) | 0.025 (df = 892) | 0.025 (df = 893) |
| *Note:* | | | *p<0.1; **p<0.05; ***p<0.01 |

Table 10: 2SLS Regression with and without ivreg Package

| | *Dependent variable:* | |
|---|---|---|
| | crime_rate | |
| | *manual 2SLS* | *ivreg 2SLS* |
| | (1) | (2) |
| fitted_business_crea | 0.111*** | |
| | (0.033) | |
| business_crea | | 0.111*** |
| | | (0.039) |
| log(pop) | −0.001 | −0.001 |
| | (0.0004) | (0.0005) |
| income | 0.0001 | 0.0001 |
| | (0.0001) | (0.0001) |
| com_typeC | 0.001 | 0.001 |
| | (0.001) | (0.001) |
| com_typeI | 0.001 | 0.001 |
| | (0.001) | (0.001) |
| Constant | −0.002 | −0.002 |
| | (0.002) | (0.003) |
| Observations | 899 | 899 |
| $R^2$ | 0.149 | -0.174 |
| Adjusted $R^2$ | 0.144 | -0.181 |
| Residual Std. Error (df = 893) | 0.004 | 0.005 |
| F Statistic | 31.260*** (df = 5; 893) | |
| *Note:* | | *p<0.1; **p<0.05; ***p<0.01 |

mean that, the better the shape of the economy in an area, the higher the crime rate even after recovering exogeneity using insrumental variables.

Table 11: 2SLS Regression With motorway and tax rate as Instruments

| | *Dependent variable:* |
|---|---|
| | crime_rate |
| business_crea | 0.217*** |
| | (0.051) |
| log(pop) | −0.002*** |
| | (0.001) |
| income | 0.0001 |
| | (0.0002) |
| com_typeC | 0.004** |
| | (0.002) |
| com_typeI | 0.004** |
| | (0.002) |
| Constant | −0.005 |
| | (0.004) |
| Observations | 899 |
| $R^2$ | -1.344 |
| Adjusted $R^2$ | -1.357 |
| Residual Std. Error | 0.006 (df = 893) |
| *Note:* | *p<0.1; **p<0.05; ***p<0.01 |