Final Exam SØK3001

Candidate number: 10046

May 2020

Exercise 1 (60 points)

In their paper "Mining and Local Corruption in Africa." published in the American Journal of Political Science in 2017, Knutsen et al. study whether natural resources have adverse effects on political institutions by increasing corruption. This question is partly inspired by their paper. The authors write: several cross-country studies indicate that dependence on natural resources is related to less democratic regime forms and worse governance institutions and outcomes. (for a recent survey, see Deacon, 2011). In particular, there seems to be a correlation between natural resources and corruption (e.g., Busse and Groning, 2013). This posited relationship may, for example, stem from natural resource revenues being relatively easy to control and monopolize for political elites (Boix, 2003; Bueno de Mesquita and Smith, 2009), in turn reducing incentives for politicians to provide accountability and transparency. Moreover, high-rent activities such as natural resource production increase the amount of resources available for patronage and unofficial transactions. Hence, one may expect that resource-abundant countries engender a political state that is factional or predatory and distorts the economy in the pursuit of rents" (Auty, 2001, 839).

a)

"Despite these plausible arguments, scholars increasingly question whether the cross-country correlations undergirding the political resource curse thesis reflect causal effects. This growing skepticism is related to an increased awareness of the limitations of traditional cross-country designs for drawing inferences." Referring to the relevant models learned in class, discuss why cross-country studies might fail at pinning down the causal effect of resources on corruption.

The skepticism in this statement refers to the use of simple OLS (Ordinary Least Squares) models when determining casual effects. The OLS method relies on minimizing the square of the residuals. It is efficient, but this does not mean that it is necessarily the best to determine good casual effects, even when controlling for important relevant factors (see explanation at the end of the reply to 1a)).

In order for the SLR/MLR OLS to be BLUE (Best Linear Unbiased Estimator), we need the following assumptions to hold:

- 1) Linearity in the parameters
- 2) Random sampling
- 3) No perfect collinearity (or $V(x_i) \neq 0$ in SLR)
- 4) Zero conditional mean, $E(u|x) = 0$
- 5) $V(u|x) = \sigma^2$

Many of these assumptions are likely to be violated, and it will be argumented that a longitudinal approach is a better way to handle the issue in question. I will mainly focus on the last two Gauss-Markov assumptions from above, as these are the ones that are most likely to be violated in the context of the task.

The zero conditional mean assumption is likely to be violated when is likely to be violated when we have important omitted factors, measurement error in the key explanatory variables and simultaneous relationships. If we face any of these issues will will have a problem of endogeneity. In the context of the model presented in this task and looking at only one time period, we should especially consider that factors may not only be contemporaneous, but perhaps there is some kind of time-lagged effect of one of our regressors on the explanatory variable.

If the homoskedasticity (constant variance) assumption about the error term is violated, OLS will still be unbiased, but no longer efficient, and inference will therefore be invalid. High collinearity of regressors may lead to nonconstant variance, and we should therefore pay attention to more the just perfectly collinear relationships.

It should be noted that the third assumption is unlikely to be violated as we rarely have perfectly collinear relationships. However, issues may also arise when we have high (though imperfect) collinearity between our regressors.

In general, using a longitudinal data set is always better. We increase the sample size and we can take into account the tendency of data over time, for instance lagged and time-constant effects.

It should also be noted that in order to determine casual effects, it is important not to omit important factors that have a covariance with our regressor and key explanatory variables. The intuition behind this can be explained using Ballentine diagrams comparing a basic SLR with and MLR (note that this intuition can be applied to more general cases of including omitted variables):

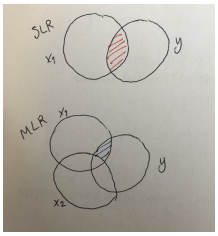

Figure 1: Ballentine diagram: Casual effects

A regression of y on x_1 in the SLR uses the full area in red to estimate β_1, while a regression of y on both x_1 and x_2 in the MLR partials out the casual effect of x_1 and x_2 and only uses the blue area.

b)

The authors adopt a different approach. They combine a novel, longitudinal data set on largescale mines, roughly for the period 1984 to 2013 with the data from Afrobarometer, a panAfrican, independent, non-partisan research network that measures public attitudes on economic, political, and social matters, over the same time period. They so obtain a panel of 33 countries. Suppose the dependent variable is the log of bribes paid by the respondent i living in neighborhood n at time t, and the independent variable is whether in the neighborhood n of the respondent i is located a mine at time t. They then estimate version of the model:

$$\log(bribe_{int}) = \beta_0 + \beta_1 mine_{nt} + \alpha_c + \alpha_t + x_0\gamma + uint, \tag{1}$$

where bribe is measured in dollars, mine is a dummy variable that takes value of one if in neighborhood n of respondent i there is a mine at time t, α_c are country fixed effects and α_t are time fixed effects and x a vector of other control variables. Explain what α_c and α_t are, and what they capture.

In model (1) we divide our data into three categories: individuals (i), neighbourhoods (n) and time (t). Our X vector captures all control variables that may vary across these categories, but there will also exist some effects that do vary across all of these, while still being important to control for. These are captured by α_c and α_t.

α_c captures all country specific effects (individuals and neighbourhoods),

meaning effects that only vary between countries but not across time. Such effects may be geography or access to other natural resources than mines.

α_c captures the time specific effects, meaning effects that does vary across time, but not across countries. These are effects are included through time indicator variables that capture "shocks" in different years. For instance, all countries were affected by the financial crisis in 2008, which is something that should be controlled for through a time indicator if one is looking at different countries within this time period.

c)

Table 1 shows results from this analysis. Interpret the results in all three columns.

The regression in the first column does not take into account any fixed effects, neither in time nor in countries. Here we estimate $\beta_1 = 0.24$, meaning that if the respondent is located at the mine, bribes are expected to increase by 24%, all else equal. This estimate seems unrealistically large and the R-squared=0.01 is very low, telling us that the estimated model has a low goodness of fit.

The second regression, with its results displayed in column (2), takes country fixed effects α_c into account, but not the time fixed effects α_t. In other words, it assumes that there are country specific fixed effects that should be accounted for, but that there have not been any time specific global occurrences that would affect our dependent variable or key regressors. Here we estimate $\beta_1 = 0.024$, meaning that if the respondent is located at the mine, bribes are expected to increase by 2,4%, all else equal. We notice that the estimated effect of mine location on bribes has decreased by a factor of ten compared to the previous model, and that the R-squared has increased substantially.

The third regression is displayed in column (3), and here all the fixed effects are included, both α_c and α_t. We now estimate $\beta_1 = 0.015$, and the standard deviation of this estimate is now very low, meaning that we have estimated with precision. The interpretation is that when all other factors remain the same, being located near mines is expected to increase you take 1,5% more money in bribary than the ones who do not have a mine in their neighbourhood.

It should be noted that all though the R-squared increases for each column, this is natural as we affect the degrees of freedom when we add more dummy variables. All though the R-squared is a measure for goodness-of-fit, we should test whether including all these fixed effects actually is relevant. A test comparing column (2) and (3) is performed in task d).

d)

A commentator suggests that during the period of interest the global economy did not incur in global crisis that could have affected the mining or the bribing activity. Evaluate this statement with an appropriate test.

This question asks us to determine whether there are exogenous time-specific effects that are important in determining our model, or if the commentator is right about there not being events in the global economy that should be controlled for.

In order to answer this question, I will test whether the inclusion of all time dummies from column (2) to column (3) is statistically significant. This is an F-test using the R-squared from column (2) and (3), where (2) is the restricted model and (3) is the unrestricted model. Using this test relies on the assumption that we have normality. We are performing the following test:

$$\log(bribe_{int}) = \beta_0 + \beta_1 mine_{nt} + \delta_0\alpha_c + \delta_1\alpha_t + x_0\gamma + uint,$$

$$H_0 : \delta_1 = 0$$
$$H_1 : \delta_1 \neq 0$$

The F-statistic takes the form:

$$F - stat = \frac{(R_u^2 - R_r^2)/q}{(1 - R_u^2)/n - k - 1}$$

In order to perform this test I will make the assumption that we have the following number of restrictions on the model: q=2013-1984-1=28. This implies a time dummy for each year of sampling, except the reference group. We cannot be sure of how many control variables we have, as we do not know the nature of the additional x-s and country fixed effects, but I assume k=35. We then compute the F-stat and get the following:

$$F - stat = \frac{(0.096 - 0.077)/28}{(1 - 0.096)/92,762 - 35 - 1} = 69,6$$

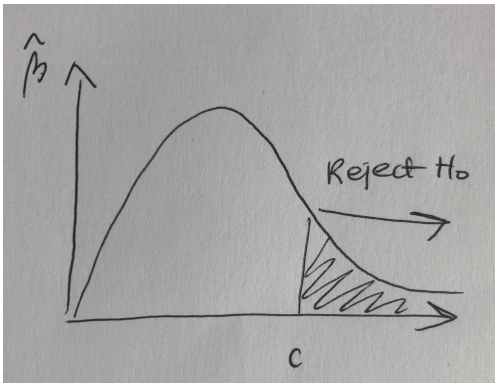

Figure 2: Rejection region

If the F-stat $> c$, we have evidence to support H_1 over H_0. Since we have an F-stat of 69,6, this exceeds all possible critical values from statisticl tables, and we have strong evidence to support rejecting the null hypothesis. The model shows strong evidence of relevant time-fixed effects, and the commentator is thereby not right that these are irrelevant.

e)

A commentator is worried that neighborhoods within countries where mines are present are substantially different than neighborhoods where mining activity is absent. In addition, these neighborhoods might be subject to very specific developments over time. Does the model in equation 1 account for such differences? If not, how could it be modified to do that?

The model in equation 1 does not account for such individual effects or neighbourhood effects. It uses the country specific effects, but this simply lumps all individuals and neighbourhoods from the same country into the same fixed effects categories, and will not capture variation within individuals and neighbourhoods.

If the model were to take into account such effects, we would need to include a set of individual fixed effects α_i and neighbourhood fixed effects α_n. However, this would raise issues when also controlling for country specific effects, as we are creating so many specific variables that we may not have sufficient data to find estimates on. To illustrate: If we had created specific effects dummies for all observations, we would have no ability to estimate the effects of such effects because we would have enough variation to use.

f)

A commentator is worried that people might misreport the amount of bribes paid, either due to social stigma or to so called recall bias. Would you be concerned about this?

Misreporting of the amount of bribes paid seems likely to occur. In a statistical context, this means that we will have a measurement error ϵ_{int} in the bribes variable, where the observed value deviates from the true population value to to some particular reason.

$$\epsilon_{int} = bribes_{int}^{observed} - bribes_{int}^{population}$$

This means that we get an error term consisting of not only the u_{int} which is iid, but also the measurement error ϵ_{int}:

$$\text{Error term} = u_{int} + \epsilon$$

We should therefor be concerned of a violation of the homoskedasticity assumption, as the variance of the error term now also will depend on the measurement error, $V(\epsilon_{int} + u_{int})$. It can be shown that:

$$V(\hat{\beta}_1) = \frac{\sigma_u^2 + \sigma_\epsilon^2}{SST_x}$$

However, measurement error in the dependent variable is not as critical as measurement error in the independent variable, as this would additionally affect the zero conditional mean assumption.

Question 2 (40 points)

One of the stylized facts in the field of political economy is that central bank independence causes comparatively lower inflation than central bank dependence. However, why that occurs is less well understood. One claim is that, when central banks are not insulated from political pressures, prime ministers and their parties manipulate monetary policy in response to changes in public opinion, especially in response to the public's evaluations of party leaders and of their expression of vote intentions. If this claim is true, then monetary variables should have no relationship with public opinion when central banks are independent. Britain is a good case in which to test this claim because the Bank of England became independent when Labor took power in mid-1997, the country's form of democracy is known for its clarity of responsibility, and it was not constrained by the European monetary system. The analysis is here performed on monthly data of two key variables for the period 1997-2006. Imagine you estimate the following model:

$$i_t = \beta_0 + \beta_1 PM_t + \beta_2 PM_{t1} + \beta_3 \log(GDP_t) + \delta t + u_t \tag{2}$$

where the variable PM_t measures the percent of respondents in the Gallup Opinion Survey who are satisfied with the performance of the prime minister; it is the monthly average short-term interest rate used for domestic monetary policy, $\log(GDP_t)$ is the (log) gross domestic product, and t represents a time trend. Table 2 shows the results.

a)

Interpret all the coefficients in column (1).

(Note regarding all use of time lags in this analysis: I have had struggles expressing lagged variables "t-1" in the document, as they seem to appear as "t1". I hope this does not lead to confusion.)

Column (1) shows the basic time series regression on model (2), which contains lags of the PM_t-regressor and a linear time trend t. It tells us the following:

$\beta_1 = -0.0339$: The monthly average short-term interest rate is expected to immediately decrease by -0.0339 when the respondents who are satisfied with the prime minister increases by one percent, all else equal.

$\beta_2 = -0.0097$: The monthly average short-term interest rate is expected to have a decrease of -0.0097 when the respondents who are satisfied with the prime minister increased by one percent in the previous time period, all else equal. We notice that the lagged effect of this key regressor is estimated to be weaker than the immediate (contemporaneous) effect of it.

$\beta_3 = 0.5732$: The monthly average short-term interest rate is expected to increase by 0.057% when GDP increases by one percent, all else equal.

$\delta = 0.0234$: The monthly average short-term interest rate is expected to increase by 0.0234 on a monthly rate, all else equal.

b)

Compute the long-run elasticity of the interest rate to public opinions. Next, explain how you would test whether it is statistically significant or not and, if possible, perform the test.

The long-run propensity (LRP) of the interest rate to public opinions is the sum of the estimated coefficients for the PM_t variable and its lag PM_{t-1}. I captures the long-run effect of the opinion about the prime minister on the monthly interest rate set by the government.

$$LRP = \beta_1 + \beta_2 = -0.0339 - 0.0097 = -0.0436$$

However, in this task we are asked to compute the long-run elasticity (LRE) for this relationship. This relies on all of the variables in question being in logarithmic forms:

$$\log(i_t) = \beta_0 + \beta_1 \log(PM_t) + \beta_2 \log(PM_{t1}) + \beta_3 \log(GDP_t) + \delta t + u_t$$

As we do not have data on such a regression, we cannot estimate the long-run propensity. If we did, we would compute it similarly to the LRP:

$$LRE = \beta_1 + \beta_2$$

I will now explain how to test the statistical significance of the long-run elasticity, meaning a test with the following hypothesis:

$$H_0 : \beta_1 + \beta_2 = 0$$
$$H_1 : \beta_1 + \beta_2 \neq 0$$

Such a hypothesis cannot be tested directly, as we would not be able to estimate the standard errors. We therefor need to reparametrize the model,

which we do not have the data to do here. However, the intuition behind it is that we create a variable $\theta = \beta_1 + \beta_2$ and we substitute for either β_1 or β_2 into the equation, and then perform a test on θ.

c)

Explain the consequences of having serially correlated errors and test whether this problem is present in the data.

Serially correlated error terms is an important concern in time series models, as this would violate the TS.1'-assumption of stationarity and weak dependence. When the error terms are serially correlated, they depend on their own previous values:

$$u_t = \rho u_{t-1} + e_t$$

Where we are in the case of a unit root if $\rho =$.

We can test for the presence of serial correlation of error terms with Breusch-Godfrey test, with the following hypothesis:

$$H_0 : \rho = 0$$
$$H_1 = \rho \neq 0$$

We use the following LM-statistic:

$$\text{LM-stat} = (n - q)R^2$$

Where n is the sample size, q is the number of restrictions (how many lags of the error term we impose), and R^2 is the R-squared from the regression of u_t on all lags. I will perform a test where the lags only last for one time-period due to the information provided in this exam, but one should ideally test for further lags, especially considering that this is a case of data being measured as frequently as every month.

$$\text{LM-stat} = (108 - 1)0.077 = 8,239$$

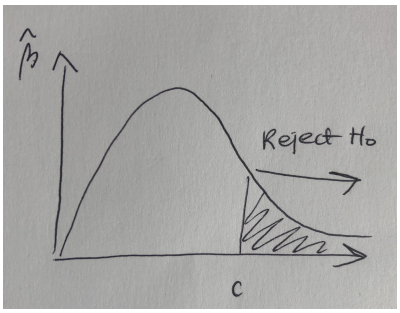

Figure 3: Rejection region

The rejection region is LM-stat ¿ c. For one degree of freedom, the critical values of the chi-square distribution are the following: 2.71 at a 10% significance level, 3.84 at a 5% significance level, and 6.63 at a 1% significance level. As 8,239 is larger than all these critical values, we have good evidence suggesting to reject the null hypothesis, and conclude that we are likely facing serially correlated error terms.

This means that our model is likely to violate one of the key assumptions for time series models to be unbiased and consistent, and our data will not be stationary around the true population value.

d)

Using the information available in the table and any other arguments you think is appropriate, discuss whether you believe that the Bank of England makes independent decisions. Propose alternative or additional

The question asks us to discuss whether we think that the Bank of England makes independent decisions. In the context of this exam, is a question of whether there is a statistically significant effect of the opinion on the prime minister on monthly interest rate set by the government.

All though our model at first sight seems to suggest such a relationship, as the PM_t is statistically significant in model (2), we are still facing issues of serially correlated error terms. This is a major concern and we should not conclude on the relationship between these key variables without controlling for it, for instance by transforming the data using a GLS transformation to achieve the following model:

$$\tilde{i_t} = \tilde{\beta_0} + \beta_1 P\tilde{M}_t + \beta_2 P\tilde{M}_{t1} + \beta_3 \log(\tilde{GDP}_t) + \delta\tilde{t} + e_t$$

The reason for these serially correlated error terms could be that we are omitting further lags of the PM_t variable which we should have included, or that there are other important factors with lagged effects which we should have

included in our model. This further implies that our estimates are likely to suffer from omitted variable bias, meaning that the effect of out potentially excluded variables may be captured by the ones we have included.

A further issue with trusting this model could be measurement error in the dependent variable. The peoples opinion on the prime minister is found using an internet survey, and it would be likely to assume that people do not always report their true opinion, or that they do not interpret the question the way they are intended to. In the case of measurement error in the independent variable, we will not only affect the variance, but the zero conditional mean assumption will also most likely be violated. This is a more serious issue than the one imposed in Exercise 1, where we discussed measurement error in the dependent variable, and will ensure that our model no longer delivers BLUE estimates. In such a case, we face endogeneity, which should be solved for using an Instrumental Variable. A suggestion for further research is therefore to use an instrumental variable for the opinion on the prime minister, which is relevant for the opinion and not correlated with our other regressors.

The decisionmaking of the Central Bank may be dependent on the people's opinion about the prime minister, but we should not use a model as simple as model (2) to conclude with this.