Econometric Methods I
Problem set 4

Samuele Bettuelli, Florian Burnat, Bella Eiriksson & Athanasios Kolokythas

November 2, 2023

This problem set includes questions. Please explain your answers clearly and state additional assumptions clearly if you use them. You may work in small groups (max 4) and you hand in 1 assignment per group.

Your solutions need to be emailed to Vincent (`vincent.pallud@bse.eu`) before 9:30 on 02-11-2023 (next Thursday). Good luck !!!

Before answering the questions below it is recommended to read the relevant pages in Hayashi (sections 1.1-1.6). See the detailed reading list on Google Classroom.

1. The Effect of the TseTse Fly on African Development

   This question is based on the paper "The Effect of the TseTse Fly on African Development" by Marcella Alsan, which was published in the American Economic Review. The paper is uploaded on Classroom and is easy and fun to read.

   The abstract of the paper is:

   The TseTse fly is unique to Africa and transmits a parasite harmful to humans and lethal to livestock. This paper tests the hypothesis that the TseTse reduced the ability of Africans to generate an agricultural surplus historically. Ethnic groups inhabiting TseTse-suitable areas were less likely to use domesticated animals and the plow, less likely to be politically centralized, and had a lower population density. These correlations are not found in the tropics outside of Africa, where the fly does not exist. The evidence suggests current economic performance is affected by the TseTse through the channel of precolonial political centralization.

   The main idea that "deep" historical roots have lasting impacts on current variation in economic development has become very popular in recent years in economics. This is sort of a local historical approach to explaining why some parts from the world are rich and others are poor can be contrasted with "general" theories of economic growth, such as the one that was empirically examined in a previous problem set.

   The data concerns several outcomes for different ethnic groups that are collected by anthropologists during the 19th century and first half of the 20th century. We are interested in the effect of the TseTse on several variables. In particular, in the file TseTse1, we have the following dependent variables (in bold are the names of the variables).

   - **animals** : large domesticated animals

- **intensive** : Intensive agriculture
- **plow** : Plow use
- **female_ag** : Female participation in agriculture
- **slavery_indigenous** : Indigenous slavery
- **central** : Centralization

These will be the y_i variables for all regressions below. To be specific y_i refers to a particular outcome for ethnic group i. The paper gives more details for all these variables.

Next, our main explanatory variable $x_{2,i}$ is a measure for the presence of the TseTse.

This is given by:

- **tse** : TseTse suitability index (intuitively: this measures the likelihood that the TseTse fly lives in the area of the particular ethnic group)

Finally, we have several control variables $x_{3,i}, x_{3,i}, \ldots, x_{12,i}$ given by:

- *Climate controls*: **meanrh** refers to humidity, **meantemp** refers to temperature, **itx** interaction between humidity and temperature
- *Malaria controls*: **malaria_index** refers to a malaria index
- *Waterway controls*: **river** refers to whether a river was located at the boundary of the ethnic group, **coast** refers to whether a coast was located at the boundary of the ethnic group
- *Geography controls*: **lon** refers to the longitude of the ethnic group, **abslat** refers to the absolute latitude of the ethnic group, **meanalt** refers to the mean altitude of the ethnic group, **SI** refers to an agricultural sustainability index

Given these variables, please answer the questions below and explain your answers clearly.

(a) Consider the regression

$$y_i = \beta_1 + x_{2,i}\beta_2 + \epsilon_i,$$

Run this regression for all six outcome variables described above. Present your results in a table. You need to show OLS estimates for β_2, s^2 and the standard errors $SE(b_2) = \sqrt{s^2((X'X)^{-1})_{22}}$.[1]

Table 1: Regression Results

	Dependent variable	$\hat{\beta}_2$	s^2	$SE(\hat{\beta}_2)$
$x_{2,1}$	animals	−0.2216	0.1868	0.0198
$x_{2,2}$	intensive	−0.2058	0.1767	0.0193
$x_{2,3}$	plow	−0.0960	0.0618	0.0114
$x_{2,4}$	female_ag	0.1255	0.2349	0.0267
$x_{2,5}$	slavery_indigenous	0.1037	0.1198	0.0171
$x_{2,6}$	central	−0.1165	0.2101	0.0214

[1] As you will see there are missing values, as is common in many studies. You can simply drop these observations.

(b) Propose a test for testing $H_0 : \beta_2 = 0$ against $\beta_2 \neq 0$, implement the testing procedure and discuss whether you accept/reject the null hypothesis for each regression.

To test the significance of β_2, we formulate the following hypothesis:

$$H_0 : \beta_2 = 0 \quad \text{(The TseTse fly has no effect)}$$
$$H_1 : \beta_2 \neq 0 \quad \text{(The TseTse fly has an effect)}$$

The testing procedure involves calculating the t-statistic, which is given by:

$$t = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)}$$

where $\hat{\beta}_2$ is the estimated value of $\beta_2$and $SE(\hat{\beta}_2)$ is the standard error of $\hat{\beta}_2$.

The results of the hypothesis tests for each dependent variable are summarized in Table 2:

Table 2: T-Test Results

	Dependent variable	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$	t	p-value	Decision
$x_{2,1}$	animals	-0.2216	0.0198	-11.1681	< 0.0001	Reject H_0
$x_{2,2}$	intensive	-0.2058	0.0193	-10.6638	< 0.0001	Reject H_0
$x_{2,3}$	plow	-0.0960	0.0114	-8.4134	< 0.0001	Reject H_0
$x_{2,4}$	female_ag	0.1255	0.0267	4.7020	< 0.0001	Reject H_0
$x_{2,5}$	slavery_indigenous	0.1037	0.0171	6.0787	< 0.0001	Reject H_0
$x_{2,6}$	central	-0.1165	0.0214	-5.4382	< 0.0001	Reject H_0

In all cases, the p-value is less than 0.001, leading us to reject the null hypothesis at any conventional significance level. This suggests that the TseTse fly suitability index has a statistically significant effect on all the considered dependent variables, indicating that areas more suitable for the TseTse fly are likely to have different socio-economic and agricultural characteristics.

(c) Now consider the regression

$$y_i = \beta_1 + x_{2,i}\beta_2 + \sum_{j=3}^{13} x_{j,i}\beta_j + \epsilon_i,$$

Run this regression for all six outcome variables and add to your table the estimates for β_2 with again s^2 and standard errors $SE(b_2) = \sqrt{s^2((X'X)^{-1})_{22}}$. Now also add the t-statistics for testing $H_0 : \beta_2 = 0$ and p-values (you do not need to present the results for the control variables).

Table 3: Regression Results with Extended Model

	Dependent variable	$\hat{\beta}_2$	s^2	$SE(\hat{\beta}_2)$	$\hat{\beta}_2$ extended	s^2 extended	$SE(\hat{\beta}_2)$ extended	t extended	p-value extended
1	animals	-0.2216	0.1868	0.0198	-0.2315	0.1689	0.0387	-5.9767	< 0.0001
2	intensive	-0.2058	0.1767	0.0193	-0.0985	0.1651	0.0383	-2.5714	0.0104
3	plow	-0.0960	0.0618	0.0114	-0.0668	0.0418	0.0193	-3.4680	0.0006
4	female_ag	0.1255	0.2349	0.0267	0.1989	0.1868	0.0506	3.9324	0.0001
5	slavery_indigenous	0.1037	0.1198	0.0171	0.1098	0.1115	0.0328	3.3485	0.0009
6	central	-0.1165	0.2101	0.0214	-0.0756	0.1967	0.0423	-1.7871	0.0746

(d) Discuss the differences between the estimates for β_2 for the models with control and without control variables.

The introduction of control variables in the regression models has led to noticeable differences in the estimates for β_2, which represents the impact of TseTse suitability on the dependent variables.

- For `animals` the introduction of the controls leads to an increase in $\hat{\beta}_2$ of 0.0099, but the relationship remains negative. This indicates Tse Tse suitability decreases with the presence of large animals.
- For `intensive` and `plow`, the introduction of control variables has led to a decrease in the magnitude of $\hat{\beta}_2$, retaining the negative estimate, indicating that the relationship between TseTse suitability and these variables is underestimated when not accounting for other factors. The decreases were -0.1073 for `intensive` and -0.0292 for `plow`
- In contrast, for `female_ag` and `slavery_indigenous`, the introduction of control variables has led to an increase in the magnitude of $\hat{\beta}_2$, maintaining the positive relationship.
- For `central`, the magnitude of $\hat{\beta}_2$ has decreased, but it remains negative, indicating that political centralization is still negatively related to TseTse suitability when other factors are taken into account, though the relationship is weaker.

These differences highlight the importance of accounting for various control variables in understanding the true relationship between TseTse suitability and different socio-economic and agricultural outcomes, ensuring that the observed relationships are not confounded by other factors.

(e) Discuss the differences between the estimates for s^2.

The inclusion of control variables in the regression models has also impacted the estimates for s^2, the variance of the error term.

- For `animals`, `intensive`, `female_ag`, `slavery_indigenous`, and `central`, the inclusion of control variables has led to a decrease in the estimate for s^2. This indicates that the model with control variables is able to better explain the variability in these dependent variables, leading to a smaller unexplained variance.

- For `plow`, the estimate for s^2 has also decreased with the inclusion of control variables, indicating a better fit of the model to the data.

In general, the decrease in s^2 across all dependent variables suggests that the extended model, which includes a variety of control variables, is able to capture more of the variability in the dependent variables, providing a better fit to the data. This highlights the importance of including relevant control variables in regression models to improve the accuracy and reliability of the results. The decrease in unexplained variance also implies that the model with control variables is more informative and provides a more comprehensive understanding of the relationships between TseTse suitability and the socio-economic and agricultural outcomes under investigation.

(f) Explain for each category of control variables (Climate, Malaria, Waterway and Geography) why these need to be included in the regression. What is the consequence of not including them?

Including control variables in a regression model allows to account for other factors that might influence the dependent variable, helping to isolate the effect of the main variable of interest. Since all of these controls are related in some way to our observations, omission of these can hold the consequence of having a dependent error term. In the context of studying the impact of the TseTse fly on various socio-economic and agricultural outcomes, the inclusion of control variables is also important for the following reasons:

- **Climate Controls**
  - *Humidity (`meanrh`)*, *Temperature (`meantemp`)*, and *Interaction between humidity and temperature (`itx`)*: Climate plays a crucial role in agricultural productivity and the spread of diseases. By controlling for humidity, temperature, and their interaction, we ensure that the effects attributed to the TseTse fly are not confounded by these climatic factors.
- **Malaria Controls**
  - *Malaria Index (`malaria_index`)*: Similar to the TseTse fly, malaria is prevalent in certain climatic conditions and can have a significant impact on health and agricultural productivity. Including a malaria index as a control helps to separate the effects of the TseTse fly from those of malaria.
- **Waterway Controls**
  - *River (`river`)* and *Coast (`coast`)*: The presence of water bodies can influence agricultural practices, trade, and development. Including variables that indicate the proximity to rivers and coasts helps to isolate the TseTse fly's effects from the broader geographic and economic impacts of being near water.
- **Geography Controls**
  - *Longitude (`lon`)*, *Absolute Latitude (`abslat`)*, *Mean Altitude (`meanalt`)*, and *Agricultural Sustainability Index (`si`)*: These geographic factors can have a profound impact on the suitability of land for agriculture, the types of crops that can be grown, and the overall development of a region.

(g) Plot the residuals e_i against $x_{2,i}$ for each regression. What can you say about the spherical error assumption?

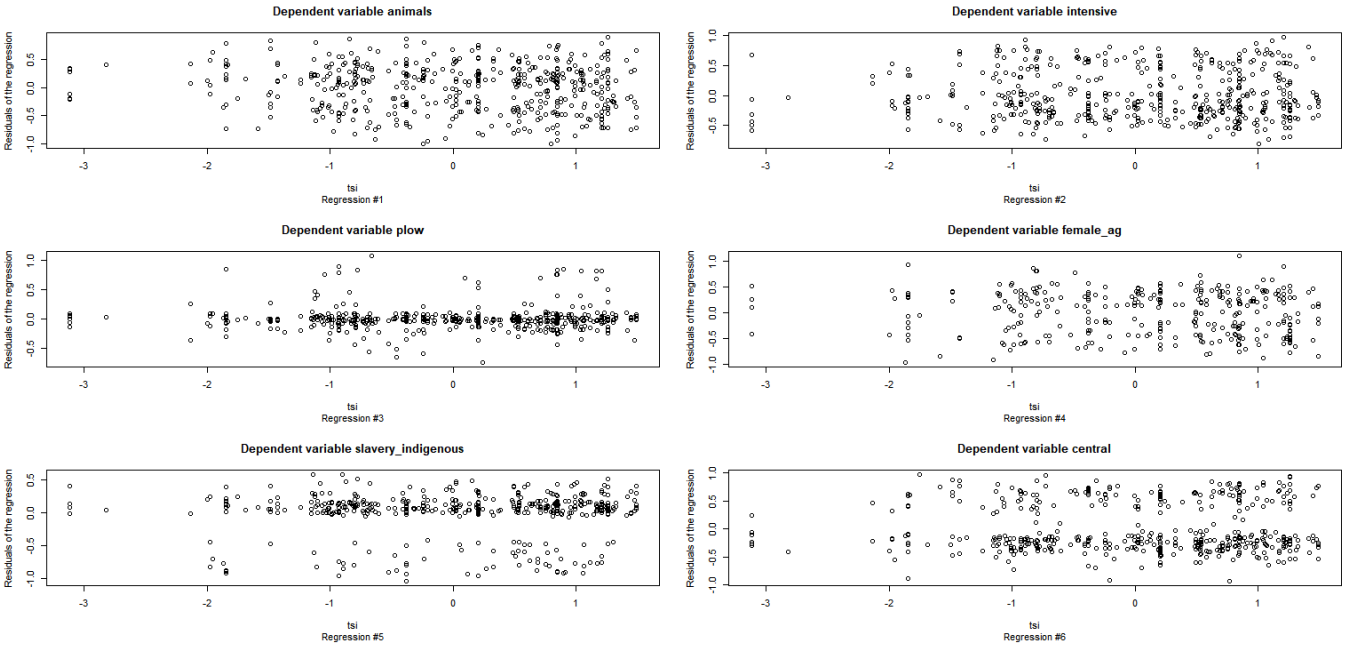

Figure 1: Plot of residuals of the regressions and the `tsi` variable

To examine whether the spherical error variance holds we need to check whether the no serial correlation and whether the homoskedasticity assumption (constant variance) hypothesis hold.

In regressions 1, 2 and 4 we observe a wide cluster of points around the horizontal line of zero and in regressions 3, 5 and 6 we observe a large concentration of points also around the horizontal line of zero, but there the points are much more concentrated than in the other three regressions.

In either case we do not observe any pattern of serial correlation between points (for example a pattern of points "going" up or down). That means that there is no serial correlation.

Homoskedasticity would appear in our plots as a random scatter of points around the horizontal line at zero. The spread of the residuals would be roughly equal across all values of x. That appears to be the case for all regressions, except for 5 and 6, where there is a non negligible part of the observations that have been spread differently than the others. That implies that the spherical error assumption holds for all regression, except of 5 (`slavery_indigenous`) and 6 (`central`).

(h) Next, a concern is that TSI is identifying a generic relationship between climate and agriculture. That means that in areas where the TseTse fly is present the climate is such that agriculture is less possible. To find out whether this is the case we use the fact that the TseTse fly only exists in Africa. In particular, we add several other ethnic groups from tropical areas from outside Africa to the dataset (NEW dataset: TseTse2) and consider the regression

$$y_i = \beta_1 + \beta_2 x_{2,i} + \sum_{j=3}^{12} x_{j,i}\beta_j + \gamma_1 A_i + \gamma_2 x_{2,i} A_i + e_i,$$

where $A_i = 1$ if the ethnic group is in Africa (**africa** in TseTse2) and zero else. Run the regression for all outcome variables and present the results for β_2 and γ_2 in a table with standard errors, t-statistics ($H_0 : \beta_2 = 0$ and $H_0 : \gamma_2 = 0$) and accompanying p-values.

Table 4: Regression Results with Africa Indicator

	Dependent variable	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$	$t_{\hat{\beta}_2}$	$p_{\hat{\beta}_2}$	$\hat{\gamma}_2$	$SE(\hat{\gamma}_2)$	$t_{\hat{\gamma}_2}$	$p_{\hat{\gamma}_2}$
$x_{2,1}$	animals	-0.0409	0.0418	-0.9793	0.3278	-0.123	0.0424	-2.895	0.0039
$x_{2,2}$	intensive	-0.0057	0.0393	-0.1450	0.8847	-0.1258	0.04	-3.148	0.0017
$x_{2,3}$	plow	-0.0164	0.0242	-0.6766	0.4989	0.0134	0.0246	0.545	0.5860
$x_{2,4}$	female_ag	0.0017	0.0462	0.0368	0.9707	0.1908	0.048	3.979	0.0000
$x_{2,5}$	slavery_indigenous	-0.0101	0.0380	-0.2646	0.7914	0.0586	0.0398	1.471	0.1417
$x_{2,6}$	central	-0.0581	0.0433	-1.3411	0.1804	-0.022	0.044	-0.502	0.6156

(i) Explain the interpretation of the coefficients for β_2 and γ_2.

- $\hat{\beta}_2$: This coefficient represents the estimated effect of the variable $x_{2,i}$ (TSI, the TseTse fly suitability index) on the dependent variable y_i across all ethnic groups outside Africa, after controlling for other variables in the model. In other words, $\hat{\beta}_2$ captures the generic relationship between the climate suitability for TseTse flies and the agricultural practices or outcomes. A significant $\hat{\beta}_2$ would suggest that climate suitability variables have an effect on agriculture variables that is common across regions included in the study. (We expect to find non-significant results, as we suppose that it's the tsetse fly that has an effect, not the climate variables linked to it)
- $\hat{\gamma}_2$: This coefficient captures the interaction effect between TSI and the Africa dummy variable A_i. Since the tsetse fly can be only found in Africa, only when A = 1 the TSI variable can actually be interpreted as the likelihood of finding the TseTse fly in the area, otherwise TSI is just a climate suitability index. It estimates how the relationship between TSI and the dependent variable y_i differs for ethnic groups in Africa compared to those outside Africa. A significant $\hat{\gamma}_2$ would indicate that the effect of TSI on agriculture is significantly different in Africa than in other tropical regions, which should imply that it is the effect of the TseTse fly that is affecting the dependent variables. Specifically, if $\hat{\gamma}_2$ is positive and significant, it would suggest that the presence of the tsetse fly in Africa increases the likelihood of using the studied agricultural techniques/observing the studied institutional behaviours among ethnic groups in Africa, the opposite would be true in the case of negative coefficients (generally the expected result).

(j) Compute the F-statistic for testing null hypothesis $\beta_2+\gamma_2 = 0$. Test the null hypothesis $\beta_2+\gamma_2 = 0$ for all outcomes using the F-test, which degrees of freedom did you use? Do these results support the claim that the TseTse fly had an effect on agricultural development?

The F-statistic is computed using the formula:

$$F = \frac{\frac{SSE_R - SSE_U}{q}}{\frac{SSE_U}{n - k_U}}$$

where:

- SSE_R is the sum of squared residuals from the restricted model (the model under H_0).
- SSE_U is the sum of squared residuals from the unrestricted model (the model without the constraint $\beta_2 + \gamma_2 = 0$).
- q is the number of restrictions, which is 1 in this case since we are only testing $\beta_2 + \gamma_2 = 0$.
- n is the number of observations.
- k_U is the number of parameters in the unrestricted model.

The degrees of freedom for the numerator is q, and for the denominator, it is $n - k_U$.

The p-value is calculated as:

$$p = 1 - F_{q,n-k_U}(F)$$

where $F_{q,n-k_U}$ is the cumulative distribution function of the F-distribution with q and $n - k_U$ degrees of freedom.

In our case, we perform this test for each dependent variable, computing the F-statistic, the p-value, and identifying the degrees of freedom used for each test.

Table 5: F-Test Results for Null Hypothesis $\beta_2 + \gamma_2 = 0$

	Dependent variable	F-statistic	p-value	df_1 numerator (q)	df_2 denominator ($n - k_U$)
1	`animals`	20.771	< 0.0001	1	729
2	`intensive`	14.974	0.0001	1	729
3	`plow`	0.021	0.8863	1	729
4	`female_ag`	20.543	< 0.0001	1	729
5	`slavery_indigenous`	2.071	0.1506	1	729
6	`central`	4.926	0.0268	1	729

- For the ”animal”, ”intensive” and ”female” regressions we found p-values that suggest to reject the null hypothesis for all the commonly used confidence levels, while for ”central” regression only at confidence level 0.05. The F-test suggests to instead accept the null hypothesis in the remaining regressions.

- Since the hypothesis is testing the significance of the total effect of TSI and its interaction with the Africa dummy, from the F-test we can conclude that a higher presence of the tsetse fly, given certain climatic conditions, has a significant impact on the likelihood of observing the usage of domesticated animals and intensive agriculture, as well as observing female participation and centralisation.