Statistical Literacy

Problem Set 11

Rémi Viné
Due December 11th, 2023

1. **General understanding and underlying assumptions.**

   (a) You are interested in predicting yearly individual income of an individual and to run the following multivariate model:

   $$Income = \beta_0 + \beta_1 Education + \beta_2 Experience + \beta_3 Primary + \beta_4 Secondary + \beta_5 Higher$$

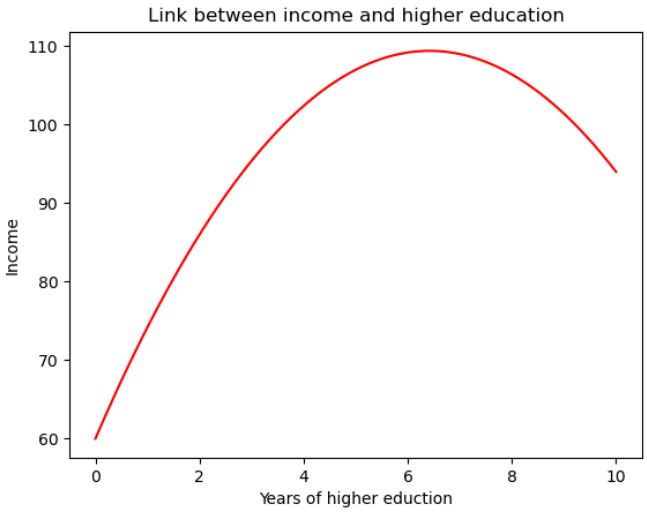
   The variable "Income" is the yearly income in thousands of Swiss Francs, the variable "Education" is the number of years of education (excluding vocational training and executive education), the variable "Experience" is the number of working years completed, the variable "Primary" is the number of years completing primary education completed, variables "Secondary" and "Higher" are, respectively, the number of years of secondary and higher education completed. Do you consider the model suitable? Explain.

   (b) You are now considering another model:

   $$Income = \beta_0 + \beta_1 Higher + \beta_2 Higher^2$$

   Do you think that this model breaks the assumption of no collinearity?

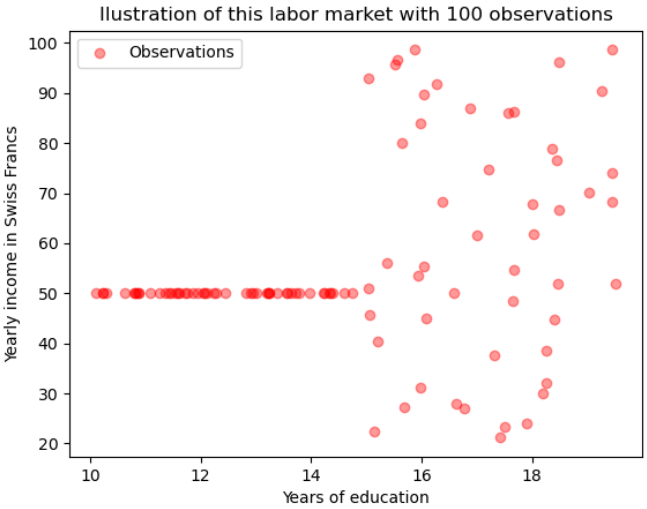
   (c) *Difficult.* Assume the regression outcome is $\hat{Income} = 60 + 15.4 Higher - 1.2 Higher^2$ (the relationship being as illustrated on the graph below). How do you interpret the relationship between years of higher education and income?



(d) Now assume that you live in a planned economy where the "labor market"" is structured as such: up to a certain level of education (15 years of schooling), every worker has to work full time and earns the same, regardless of the experience. Above this threshold, workers (once they completed their studies) can choose whether they prefer to work part time and salaries vary. Experience is not related with the income earned. Running OLS on the following regression:

$$Income = \beta_0 + \beta_1 Education$$

, and the two variables would be as represented on the scatter plot below. Would the specific labor market structure be problematic?



(e) You are running two linear models and obtain:

$$(1) \quad \hat{Income} = 94.2 + 17.0Higher$$

and

$$(2) \quad \hat{Income} = 52.7 + 16.5Higher + 5.5Experience$$

Discuss the difference in interpretation of the b_1 coefficients.

(f) Using this model:

$$\hat{Income} = 52.7 + 16.5Higher + 5.5Experience$$

, would you advise a young person (looking to maximize future income) to study or to work directly (in order to have a higher income)?

2. **Urban violence in cities.**

Assume that we are interested in urban violence events. We want to see what factors can help predict the number of deaths in a city (over a given period of time) due to civil disorders. Table 1 explains the different variables used. Tables 2 and 3 show the outcomes of a few OLS models. Until question l), refer to results in table 2.

Label	Details
Number of death events	Number of death events in a city (1960-2014)
Number of non-deadly events	Number of non-deadly events in a city (1960-2014)
(Log) GDP per capita	Country GDP per capita, log-transformed (2014)
Population in city ('000 000)	Population in a city in million inhabitants (2014)
Autocracy	Binary variable equal to one if if a country is considered autocratic
Democracy	Binary variable equal to one if a country is considered as democratic
Economic Shock	Binary variable equal to one if there is an economic shock in the country
Year	Year when the count of urban disorders was implemented
Continent binary variables	Binary variables equal to one if the city belongs to a given continent

Table 1: Codebook - simplified

(a) What is the correlation between "death events" and "Number of non-deadly events"?

(b) Are all the models referring to the same population?

(c) Using Model 1, how many deadly events are expected to occur if there is not non-deadly civil disorder happening? Find the answer to this question in model 2 and compare with your answer in model 1 (if possible).

(d) Looking at all five models, would you conclude that deadly events and non deadly events are positively or negatively related? Explain your conclusion at $\alpha = 5\%$.

(e) Looking at Model 3, is there a region of the world that is missing, and, if so, why?

(f) Looking at Model 3, what is the reference region of the world (one that is not shown on the table outputs), and, why is there a reference region?

(g) Using Model 3, how to interpret the estimator attached to the variable "Population in city ('000 000)"?

(h) Using Model 3, at $\alpha = 0.05$, would you conclude that the variable "Population in city ('000 000)" is a statistically significant predictor?

(i) Using Model 2, where the stars informing on $p - value$ (if any) are hidden on the table, run the appropriate hypothesis testing in order to know whether zero, one, two or three stars would be allocated to the estimator b_7 attached to the variable "Economic shock".

(j) Assume you run a sixth model where $R^2 = 0.869$ and $Adjusted\ R^2 = 0.356$. What can you say about the fit of the model and the (statistical) usefulness of all included variables?

(k) Using Table 3, what type of model selection was implemented from "Model 2" to "Model 2 ter"?

(l) Using Table 3 and using the model whose $p - value$ for the F-test is the *lowest*, would you conclude, *ceteris paribus*, that there are more deadly events in democratic regimes or in autocratic regimes?

Table 2: OLS Models

	Dependent variable, Y: Number of death events				
	Model 1	Model 2	Model 3	Model 4 (Year: 2015)	Model 5 (Continent: SSA)
	(1)	(2)	(3)	(4)	(5)
Number of non-deadly events	0.260***	0.233***	0.229***	0.342***	0.304***
	(0.008)	(0.008)	(0.008)	(0.062)	(0.019)
(Log) GDP per capita		-0.029**	-0.108***	-0.075	-0.143***
		(0.015)	(0.019)	(0.169)	(0.026)
Population in city ('000 000)		0.005	0.007*	-0.028	0.030**
		(0.004)	(0.004)	(0.022)	(0.014)
Autocracy		-0.187***	-0.274***	-0.221	-0.206***
		(0.042)	(0.042)	(0.464)	(0.044)
Civil conflict		0.733***	0.679***	1.606***	0.333***
		(0.040)	(0.040)	(0.334)	(0.046)
Democracy		-0.199***	-0.126***	0.691**	-0.094
		(0.043)	(0.045)	(0.306)	(0.060)
Economic shock		0.120 [redacted]	0.101***	0.226	0.028
		(0.033)	(0.033)	(0.281)	(0.039)
Europe			0.031	-0.512	
			(0.061)	(0.446)	
LATAM			0.114**	-0.357	
			(0.054)	(0.434)	
MENA			0.708***	1.048**	
			(0.060)	(0.467)	
North America			0.155	0.700	
			(0.099)	(0.770)	
Oceania			0.144	-0.194	
			(0.108)	(0.842)	
SSA			-0.040	0.142	
			(0.048)	(0.426)	
Year		0.003***	0.003***		
		(0.001)	(0.001)		
const	0.236***	-5.669***	-5.328**	0.240	1.295***
	(0.017)	(2.090)	(2.130)	(1.535)	(0.203)
Observations	7,710	7,710	7,710	161	2,027
R^2	0.119	0.175	0.195	0.434	0.185
Adjusted R^2	0.119	0.174	0.194	0.384	0.182
F Statistic	1037.539*** (df = 1.0; 7708.0)	203.601*** (df = 8.0; 7701.0)	133.474*** (df = 14.0; 7695.0)	8.688*** (df = 13.0; 147.0)	65.340*** (df = 7.0; 2019.0)

Note: *p<0.1; **p<0.05; ***p<0.01
Note 2: Standard errors in parentheses

Table 3: OLS Models

	Dependent variable, Y: Number of deadly events		
	Model 2	Model 2 bis	Model 2 ter
	(1)	(2)	(3)
Number of non-deadly events	0.233***	0.234***	0.233***
	(0.008)	(0.008)	(0.008)
(Log) GDP per capita	-0.029**	-0.028*	
	(0.015)	(0.015)	
Population in city ('000 000)	0.005		
	(0.004)		
Autocracy	-0.187***	-0.186***	-0.191***
	(0.042)	(0.042)	(0.042)
Civil conflict	0.733***	0.739***	0.755***
	(0.040)	(0.040)	(0.039)
Democracy	-0.199***	-0.192***	-0.225***
	(0.043)	(0.043)	(0.039)
Economic shock	0.120***	0.117***	0.120***
	(0.033)	(0.033)	(0.033)
Year	0.003***	0.003***	0.003***
	(0.001)	(0.001)	(0.001)
const	-5.669***	-6.022***	-5.541***
	(2.090)	(2.074)	(2.059)
Observations	7,710	7,710	7,710
R^2	0.175	0.174	0.174
Adjusted R^2	0.174	0.174	0.173
F Statistic	203.601*** (df = 8.0; 7701.0)	232.397*** (df = 7.0; 7702.0)	270.447*** (df = 6.0; 7703.0)

Note: *p<0.1; **p<0.05; ***p<0.01
Note 2: Standard errors in parentheses