José António Ferreira Machado
Paulo Fagandini
Marlon Francisco
Nova SBE

Econometrics
Spring 2024-25

1 Exercises - The Simple Linear Regression Model

1.1 Europe is in the midst of an unprecedented human migration with hundreds of thousands flocking to Europe's shores. You want to study the root causes of this issue and decide to do some research.

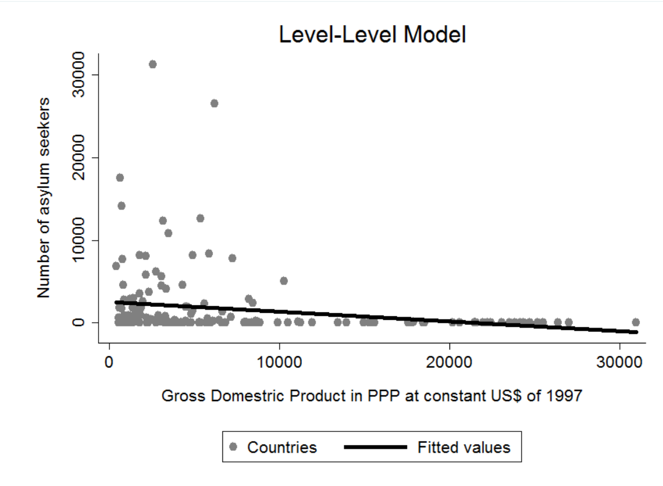
a) Write an equation that would allow you to test whether a country's GDP has a linear influence in its number of asylum seekers.

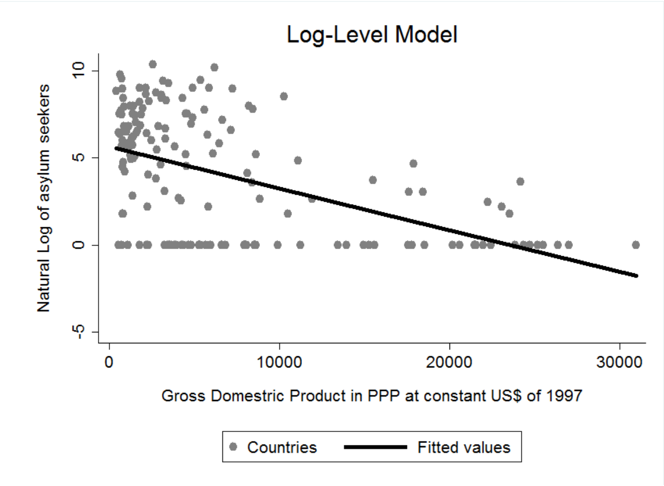
b) Can you say there is an exact linear relationship between both variables?

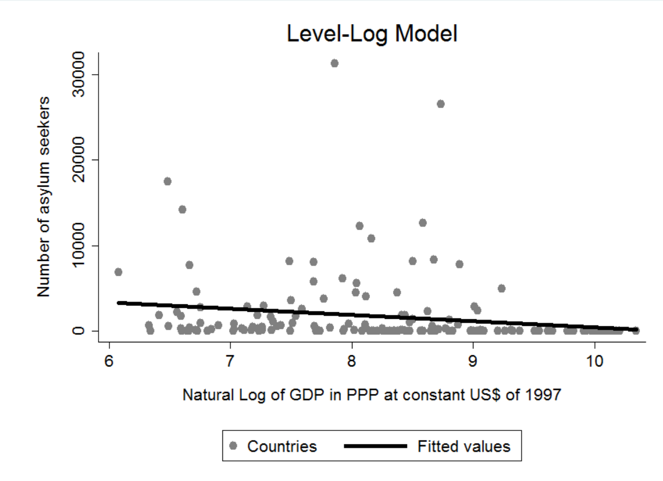
c) Can you easily compute the population parameters for the regression model you presented before?

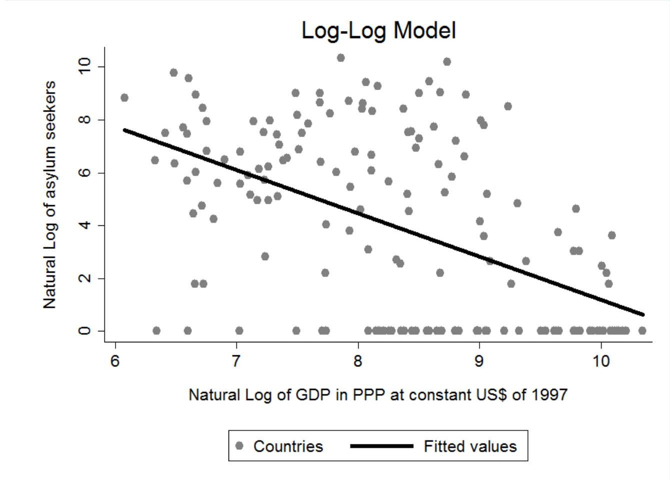
d) What do you suggest then?

1.2 You move on with your endeavour and collect a set of cross-sectional data for 166 countries on their number of asylum seekers and GDP per capita in 1999 and estimate the following models by OLS:









$$Level - Level : \widehat{asylum} = 2569 - 0.12GDP \qquad n = 166, R^2 = 0.0447 \tag{1}$$

$$Log - Level : \log(\widehat{asylum}) = 5.64 - 0.0002GDP \qquad n = 166, R^2 = 0.2585 \tag{2}$$

$$Level - Log : \widehat{asylum} = 7804 - 731.97\log(GDP) \qquad n = 166, R^2 = 0.0368 \tag{3}$$

$$Log - Log : \log(\widehat{asylum}) = 17.56 - 1.64\log(GDP) \qquad n = 166, R^2 = 0.2673 \tag{4}$$

(a) Interpret the coefficient on the regressor of each model.

(b) Taking into account this econometric exercise alone can you say that higher GDP causes the number of asylum seekers to decrease?

(c) You happen to read in a newspaper the following sentence: "The asylum migration flows are determined by economic incentives alone". Does the R^2 of any of the fitted models provide evidence in favour of this claim?

(d) What can you conclude when you compare the quality of adjustment of the four regressions?

(e) Intuitively, what do you point out as the main limitations of these models?

1.3 The following equation relates housing price (*price*) to the distance from a garbage incinerator (*dist*):

$$\log(\widehat{price}) = 9.40 + 0.312\log(dist) \qquad n = 135, R^2 = 0.162$$

(a) Interpret the coefficient on $\log(dist)$. Is the sign of this estimate what you expect it to be?

(b) Do you think simple linear regression provides an unbiased estimator of the ceteris paribus elasticity of *price* with respect to *dist*? (Think about the city's decision on where to put the incinerator.)

(c) What other factors about a house affect its price? Might these be correlated with distance from the incinerator?

1.4 Francis Galton set up the term "regression" in an influential paper published in 1886, "Regression Towards Mediocrity in Hereditary Stature" where he examined the joint distribution of the stature (height) of parents and children. Galton discovered that on average a child's height is more mediocre than his or her parent's height, namely 2/3. In the following table y is the height of children and

x is the height of parents (measured in terms of deviations from average height, 165cm). Note that although you have data in terms of deviations you should use it as it is on the table in order to estimate the parameters.

y	-2	0	1	0	1
x	-2	-1	0	1	2

a) Estimate, by OLS, the parameters $(\alpha, \beta, \gamma, \delta)$ of the following models:

$$y = \alpha + \beta x + \epsilon$$

$$x = \gamma + \delta y + \upsilon$$

b) Are the two fitted regression lines the same? Explain it recurring to the difference between correlation and regression between two variables.

1.5 Consider the savings function:

$$sav = \beta_0 + \beta_1 inc + u, u = \sqrt{inc} \cdot e,$$

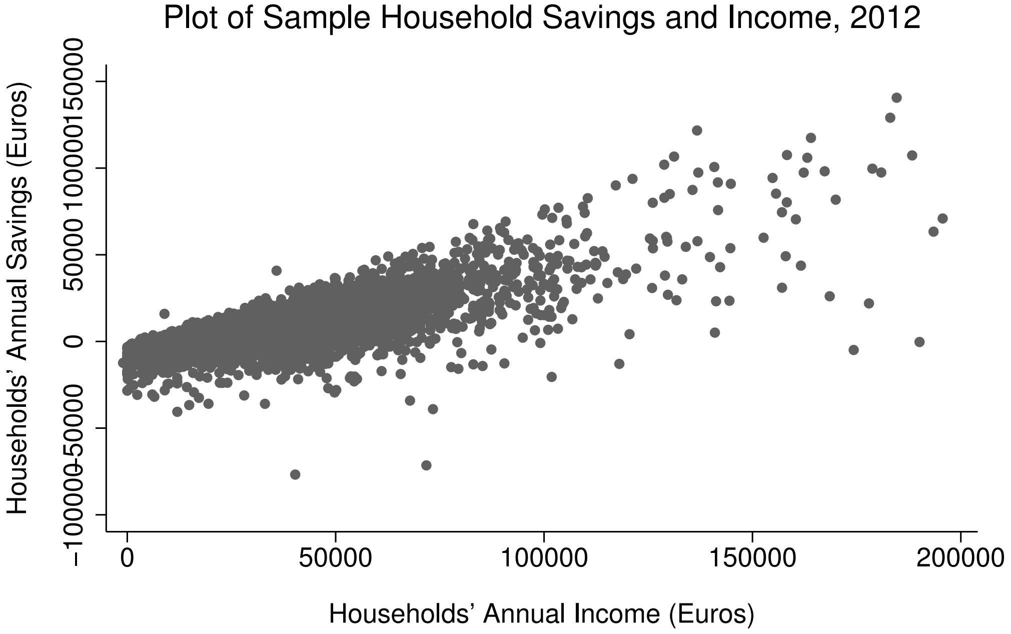
where e is a random variable with $E(e) = 0$ and $Var(e) = \sigma_e^2$. Assume that e is independent of inc.

a) Show that $E(u|inc) = 0$, so that the key zero conditional mean assumption (SLR.4) is satisfied. [Hint: If e is independent of inc, then $E(e|inc) = E(e)$].

b) Suppose now that $E(u) = \alpha \neq 0$. Show that the model can always be rewritten with the same slope, but a new intercept and error, where the new error has a zero expected value.

c) Show that $Var(u|inc) = \sigma_e^2 \cdot inc$, so that homoskedasticity (SLR.5) is violated. In particular, the variance of sav increases with inc. [Hint: $Var(e|inc) = Var(e)$, if e and inc are independent.]

d) Provide a discussion that supports the assumption that the variance of savings increases with family income. Do you find evidence of that in the graph below?



Source: Survey of Household Income and Wealth 2012, Banca d'Italia (n=8151)

1.6 The following table contains the ACT (American College Testing) scores and the GPA (grade point average) for eight college students. Grade point average is based on a four-point scale and has been rounded to one digit after the decimal.

Student	**GPA**	**ACT**
1	2.8	21
2	3.4	24
3	3.0	26
4	3.5	27
5	3.6	29
6	3.0	25
7	2.7	25
8	3.7	30

(a) Estimate the relationship between GPA and ACT using OLS; that is, obtain the intercept and slope estimates in the equation:

$$\widehat{GPA} = \hat{\beta}_0 + \hat{\beta}_1 ACT$$

Comment on the direction of the relationship. Does the intercept have a useful interpretation here? Explain. How much higher is the GPA predicted to be if the ACT score is increased by five points? (Note: the ACT is a standardized test for high school achievement and college admissions in the United States).

(b) Compute the fitted values and residuals for each observation, and verify that the residuals (approximately) sum up to zero.

(c) What is the predicted value of GPA when ACT = 20?

(d) How much of the variation in GPA for these eight students is explained by ACT? Explain.

1.7 Consider 41 monthly observations (January 1991 to May 1994) on the Portuguese Consumer Price Index without housing expenses (basis is the average of prices in 1991) and on the net public assets (in billions of escudos). Using this dataset, the following models were estimated by OLS:

$$\begin{aligned}
\widehat{CPI} &= 44.484 + 5.686L & R^2 &= 0.966 \\
\widehat{\ln(CPI)} &= 3.289 + 0.578\ln(L) & R^2 &= 0.964 \\
\widehat{\ln(CPI)} &= 4.094 + 0.052L & R^2 &= 0.973 \\
\widehat{CPI} &= -42.635 + 62.625\ln(L) & R^2 &= 0.954
\end{aligned}$$

a) Interpret the estimated parameters for each model.

b) Briefly discuss the coefficients of determination (R^2).

c) Given the results presented above, which model would you choose? Why?

1.8 The Engel Curve shows the relationship between the various quantities of a good that a consumer is wiling to purchase at varying income levels. In a survey with 40 households data were obtained on expenditure on dairy products and income.

a) Using this database (Engel.xls), obtain estimates and the respective R^2 for the following models:

$$Linear - log\ model : dairy = \beta_0 + \beta_1 \ln(inc) + u$$

$$Log - log\ model : \ln(dairy) = \beta_0 + \beta_1 \ln(inc) + u$$

b) Estimate the marginal propensity ($\partial dairy/\partial inc$) to expenditure and elasticity expenditure/income ($\partial dairy/\partial inc \times inc/dairy$) for both models. Interpret them economically.

c) Can you conclude which is the best model by looking at the coefficient of determination?

1.9 Consider the standard simple regression model $y = \beta_0 + \beta_1 x + u$ under the Gauss-Markov Assumptions SLR.1 to SLR.5. The usual OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased for their respective population parameters. Let $\tilde{\beta}_1$ be the estimator of β_1 obtained by assuming the intercept is zero.

a) Compute $\tilde{\beta}_1$ by Ordinary Least Squares.

b) Verify that $\tilde{\beta}_1$ is unbiased for β_1 when the population intercept (β_0) is zero. Are there other cases where $\tilde{\beta}_1$ is unbiased?

c) Find the variance of $\tilde{\beta}_1$.[Hint: The variance does not depend on β_0.]

d) Show that $\text{Var}(\tilde{\beta}_1) \leq \text{Var}(\hat{\beta}_1)$. [Hint: For any sample of data, $\sum_{i=1}^{n} x_i^2 \geq \sum_{x=1}^{n}(x_i - \bar{x})^2$, with strict inequality unless $\bar{x} = 0$.]

e) Comment on the trade-off between bias and variance when choosing between $\hat{\beta}_1$ and $\tilde{\beta}_1$.

1.10 Consider now that the true model for a population is $y_i = \beta_1 x_i + u_i$ and that all the classical linear model assumptions hold. Namely you can assume that x represents income and y represents income tax revenue, as illustrated in Wooldridge's book. Assume also that $Var(u_i) = \sigma^2$. One student has chosen the following estimator for β_1 (where n is the number of observations):

$$\tilde{\beta}_1 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{x_i}\right)$$

a) Prove that the estimator is unbiased, under the extra condition in a).

b) Calculate the variance of the estimator.

c) Compare $Var(\hat{\beta}_1)$ with $Var(\tilde{\beta}_1)$ where $\hat{\beta}_1$ is the estimator of β_1 in a model where the intercept is assumed to be zero. [Hint: $Var(\hat{\beta}_1)$ was found in exercise 1.9 c).]

1.11 Consider the standard SLR $y = \beta_0 + \beta_1 x + u$ under the Gauss-Markov Assumptions SLR.1 to SLR.5. The usual OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased for their respective population parameters.

a) Let c_1 and c_2, with $c_2 \neq 0$, be constants (you can assume they represent exchange rates between different currencies). Let $\tilde{\beta}_0$ and $\tilde{\beta}_1$ be the intercept and slope from the regression of $c_1 y_i$ on $c_2 x_i$. Show that $\tilde{\beta}_1 = (c_1/c_2)\hat{\beta}_1$ and $\tilde{\beta}_0 = c_1\hat{\beta}_0$. [Hint: To obtain $\tilde{\beta}_1$, plug the scaled versions of x and y into the formula of $\hat{\beta}_1$ derived for the SLR model. Use it to find $\tilde{\beta}_0$ in the rescaled formula of $\hat{\beta}_0$.]

b) Now, let $\tilde{\beta}_0$ and $\tilde{\beta}_1$ be from the estimators from the regression of $(c_1 + y_i)$ on $(c_2 + x_i)$ (with no restriction on c_1 or c_2, in here you can assume that these constants represent lump-sum taxes). Show that $\tilde{\beta}_1 = \hat{\beta}_1$ and $\tilde{\beta}_0 = \hat{\beta}_0 + c_1 - c_2\hat{\beta}_1$.

1.12 Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be the OLS intercept and slope estimators, respectively, and let $\bar{u}$ be the sample average of the errors (not the residuals!) and prove some steps of the theorems of unbiasedness and sampling variance of OLS estimators

(a) Show that $\hat{\beta}_1$ can be written as $\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} w_i u_i$ where $w_i = d_i/SST_x$ and $d_i = x_i - \bar{x}$.

(b) Use part (a), along with $\sum_{i=1}^{n} w_i = 0$ to show that $\hat{\beta}_1$ and $\bar{u}$ are uncorrelated. [Hint: You are being asked to show that $E[(\hat{\beta}_1 - \beta_1)\cdot\bar{u}] = 0$.]

(c) Show that $\hat{\beta}_0$ can be witten as $\hat{\beta}_0 = \beta_0 + \bar{u} - (\hat{\beta}_1 - \beta_1)\bar{x}$.

(d) Use questions (b) and (c) to show that $Var(\hat{\beta}_0) = \sigma^2/n + \sigma^2(\bar{x}^2)/SST_x$.

(e) Show that: $Var(\hat{\beta}_0) = \frac{\frac{\sigma^2}{n}\sum_{i=1}^{n} x_i^2}{SST_x}$. [Hint: $SST_x/n = \frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2$]