Worksheet 2
The simple linear regression model

1. The joint probability for two discrete variables can be represented by the following table:

$P(Y,X)$	$X=1$	$X=2$	$X=3$
$Y=0$	0.15	0.10	0.15
$Y=1$	0.15	0.30	0.15

   a) Get the conditional expectation function, $E(Y \mid X)$.

   b) Get the linear predictor $L(Y \mid X)$.

   c) Present a table with the values of $E(Y \mid X = x_j)$ and $L(Y \mid X = x_j)$ for the values $x_j = 1, 2, 3$.

   d) Obtain the marginal distribution for the random variables $E(Y \mid X)$ and $E(X \mid Y)$.

2. Consider the **conditional** probability distributions of Y given X, $P(Y \mid X = x_j)$, $x_j = 0, 5, 10$.

$P(Y \mid X = x_j)$		X		
		0	**5**	**10**
Y	**5**	1/3	1/2	1/3
	10	1/3	1/2	1/3
	15	1/3	0	1/3

   and the marginal probability distribution of X, $P(X)$

$P(X)$	X		
	0	**5**	**10**
	3/10	4/10	3/10

   a) Get the conditional expectation function, $E(Y \mid X)$.

   b) Get the linear predictor $L(Y \mid X)$.

   c) Present a table with the values of $E(Y \mid X = x_j)$ and $L(Y \mid X = x_j)$ for the values $x_j = 0, 5, 10$.

   d) Obtain the marginal distribution for the random variables $E(Y \mid X)$ and $E(X \mid Y)$.

3. Assume that $Y = X + U$, where

$$E(X) = 100, \quad E(U) = 0, \quad V(X) = 600,$$
$$V(U) = 1000, \quad C(X,U) = 400.$$

   You are informed that a person has a value of $Y = 110$. Predict in the best possible way his/her value of X.

4. Assume that the random variables X and Y are independent. Show that the product $E(XY)$ is equal to the product of the expected values, $E(X)E(Y)$. What is the implication of this finding on the value of $C(X,Y)$?

5. The data file `DSTAR.GDT` contains information for 5743 students who participated in the experiment for 4 consecutive years, being randomly assigned from the first year in a class of a certain size. The variables included in the dataset are *schidkn*, *classid* (student and class identifiers, respectively) *pscore* (score obtained in an objective test called the SAT) *size* (class size) *small* (which takes the value 1 if the class is small – between 13 and 17 pupils, and 0 if standard-between 22 and 25 students-) *female* (which takes the value 1 if the student is a girl and 0 otherwise) *nonwhite* (which takes the value 0 if the student is white and 1 otherwise). It should be noted that for various reasons, there are classes whose size differ from those classified as small or standard, in that case the *small* variable has no value available for those observations.

   a) Using the program Gretl, perform an analysis of variance for qualifying *pscore* variable depending on the *small*, which indicates whether the class is small or standard. Fot that purpose, execute Gretl, open `DSTAR.GDT` (File → Open Data→ User File, and choose the corresponding directory and file); then, go to Model→ Other linear models→ ANOVA and select *pscore* as "response variable" and *small* as "treatment variable".

   b) Test that the average rating is independent of class size, measured according to the variable *small*.

   c) Perform a least square regression of the rating *pscore* on the size *size*. For such purpose, execute Gretl and open `DSTAR.GDT`, go to Model→ Ordinary Least Squares, select *pscore* as "response variable" and add *size* as "treatment variable".

   d) What is your conclusion about the effect of class size on academic performance?

   e) How would your conclusion if during the experiment students had changed from class originally assigned to other of different sizes? (For example, due to the insistence of those parents whose children were initially placed in standard classes).

6. In an alternative specification of the classical linear regression, Can we replace the assumption $E(\varepsilon|x) = 0$ for the assumption $E(\varepsilon) = 0$? Are these assumptions equivalent? Is it possible that $E(\varepsilon) = 0$ and $E(\varepsilon|x) = 0$ for all x? Is it possible that $E(\varepsilon|x) = 0$ for every x but $E(\varepsilon) = 0$?

   (**Hint:** ¿Is it possible that the average income in the different states of the US were equal to 20000 dollars without the rent for every state being equal to 20000 dollars? Is it possible that the average income for every state were 20000 dollars but the average income for the US being different to 20000 dollars?).

7. Assume that in order to establish the linear relationship between Y = percentage variation in the real wages and X = unemployment rate (in %) we consider the following expression:

$$Y = 8{,}33 - 0{,}84X + u.$$

   a) Interpret the meaning of the estimated coefficients.

   b) Assume that we consider the specification with the inverse function $X' = 1/X$ (the inverse of the unemployment rate) as independent variable:

$$Y = -0{,}12 + 0{,}983X' + \varepsilon'.$$

   Interpret the meaning of the estimated coefficients.

8. The dataset `ALI.GDT` contains cross section data with 965 randomly chosen observations for couples, with or without children, in which the age of the man ranges between 25 and 65 years (Source: *Encuesta de Presupuestos Familiares 1990-91*, elaborated by the Spanish *Instituto Nacional de Estadística*). The survey contains the following variables: $V1 = AL$ = annual food expenditure of the family (in euros), $V2 = GT$ = complete annual expenditure of the family (in euros), $V3 = RF$ = annual income of the family (in euros), $V4 = NH$ = Number of children younger than 18 years of age, $V5 = NA$ = number of adults (including the spouse), $V6 = EH$ = husband's age, $V7 = EM$ = wife's age, $V8 = UH$ = university studies of the husband (taking a value of 1 in case of holding a college degree and 0 otherwise), $V9 = UM$ = university studies of the wife (taking a value of 1 in case of holding a college degree and 0 otherwise), $V10 = MT$ = labor status of the wife (taking the value of 1 in case of working and 0, otherwise). Consider the following variables:

$$\begin{aligned} V1 &= \text{Annual family expenditure on food (in euros)} \\ V2 &= \text{Total annual family expenditure (en euros)} \end{aligned}$$

We have used the same data to estimate three alternative models, whose estimated coefficients are presented in the following lines. Check the estimated coefficients with the help of Gretl, and give an interpretation of the coefficients for each of the specifications.

a) $\ln(V1) = 3{,}67 + 0{,}48\ln(V2)$;

b) $V1 = -164567 + 2163\ln(V2)$;

c) $(V1/V2) \times 100 = 156{,}89 - 13{,}32\ln(V2)$ ("Working-Lesser"specification, which consider the determinants of food expenditure as a percentage of total expenditure).

9. A multinational, with 1120 branches spread across the whole world, wishes to study the fundamentals of sales. In order to do that, the following model is proposed:

$$[\ln(V) - \ln(NH)] = \beta_0 + \beta_1 [\ln(R) - \ln(NH)] + \varepsilon,$$

where the error term ε satisfies the classical regression model assumptions and

$$\begin{aligned} V &= \text{Annual Sales (thousand of dollars) for an specific brunch,} \\ R &= \text{Aggregate disposable income (thousand of dollars) in locality where the brunch is located,} \\ NH &= \text{Population of the locality where the brunch is located.} \end{aligned}$$

Give an interpretation for β_1.

10. The *per capita* consumption of electric energy, in thousand of of kWh, and the *per capita* income, in thousand of euros for the countries belonging to the European Union the year 2001 are explained for the following linear model,

$$C = -0{,}154 + 0{,}571R + \varepsilon.$$

Compute the *per capita* income elasticity for a *per capita* income of 6000 euros.

11. The CAPM (Capital Asset Pricing Model) is an equilibrium model explaining the expected returns for assets. The regression for the excess of return (over the risk-free asset) has the following simple econometric specification:

$$(R_i - r_i^f) = \beta_1 + \beta_2(R_i^M - r_i^f) + \varepsilon_i,$$

where, for the i-th month, R_i represents the return of the asset, r_i^f is monthly return of the risk-free asset (for example, the Treasury bills with a maturity of 30 days), R_i^M is the return of the market portfolio (i.e., the average **weighted** return for a portfolio composed for **all** market available assets), and ε_i is the error term that captures the random fluctuations that are independent on the market portfolio.

a) Give an interpretation for β_2.

b) What can we say about an asset with $\beta_2 = 1$? And one with $\beta_2 > 1$? And with $\beta_2 < 1$?