Prof. Javier Gómez Biscarri
MSc in Finance/EPP/CMR
ECONOMETRICS

# Assignment 1: The first on multiple regression

**Special directions**: The first exercise can be done individually or in groups. Your answers must be submitted by mail to daniel.dejuan@barcelonagse.eu before November 5, 7 pm. The second exercise is INDIVIDUAL and the deliverable is an Excel file that you must submit **by email** to me at javier.gomez@upf.edu. Submission deadline for the second exercise is November 3, 3 pm.

You have in the Classroom an excel file named *data_problemset1.xls* with two worksheets that have two separate datasets (you will need to input those datasets into Stata). Both datasets are examples of a hedonic price analysis, where the price of a certain good is examined from the point of view of characteristics of the good: when doing this we implicitly assume that the market –or the buyer and seller in a negotiated sale- "prices" differentiated versions of a good in terms of the differential features, so in some sense it is the features that are being priced. Of course, applications of this type of analysis are numerous and in all kinds of fields. We will use in this problem set two examples of goods with which all of us are familiar: cars and houses.

1) **A traditional regression exercise on price analysis.**
The first worksheet is called "Car Prices". The last column contains the **selling price of a set of used BMW's**. The other columns contain a number of **characteristics** of the car that we believe affect (or help explain) the selling price. Most of the variables are self explanatory, except for the following:

a) Gas_1 / Diesel_0 is a 1 if the car runs on regular gas and a 0 if it uses Diesel.
b) ABS is a 1 if the car has antiblocking braking system and 0 otherwise.
c) ABD is a 1 if the car has passenger airbag and a 0 otherwise.
d) CD is a 1 if the car has a CD player and a 0 otherwise.
e) AW is a 1 if the car has alloy wheels and a 0 otherwise.
f) A/C is a 1 if the car has air conditioning and a 0 otherwise.
g) MP is a 1 if the car has metal paint and a 0 otherwise.
h) YEAR is the year of design of the car (so "1997" means an older car than "2001")

Your task is to analyze the determinants of the price of used cars in a MULTIPLE REGRESSION setting.

a) Estimate first a simple model where the only explanatory variable is *kilometers*. Interpret (i.e., explain the meaning and use of) the following output:
- Parameter values and t-statistics (significance of the two parameters).
- R2 of the model and the F-value.

b) Construct and interpret a confidence interval for the marginal effect of *kilometers*.

c) Graph the residuals against *kilometers*. Do you see any evidence of a nonlinear effect? Explain what could cause this apparent nonlinearity. Test statistically whether indeed the effect of *kilometers* is nonlinear. Interpret the conclusion of the test. How will you modify the model so that the nonlinearity is accounted for?

For the rest of the exercise, exclude all nonlinearities and use a simple linear model.

d) Select now a more complete model where you include other relevant variables (maybe a look at the correlation matrix - Stata command: **correlate** - can help you *start* the process).
- Interpret the parameters and t-stats of your final model of choice.
- Test whether the new additional variables have significant joint explanatory power (that is, do they add anything to the simpler model you estimated in c?).

e) Once you have your final model, test the following hypotheses (in some cases, you may have to include variables that you had excluded in the final model):
1. The price of the car goes down by 1200 euros per additional year old the car is.
2. The number of CC and ABS have no impact on the price of the car.
3. The effect of having a CD player is the same as the difference between a gas and a diesel car.
4. Two cars that have the same characteristics except that one has 20000 more kilometers but it is of a design 4 years newer will sell for the same price.
5. There is no heteroskedasticity problem (White's test).
6. There is no heteroskedasticity related to YEAR or HP (Breusch-Pagan test).

f) Estimate the model with and without robust standard errors. Do the robust errors –and, therefore, the conclusions on significance- differ much from regular OLS standard errors?

g) You work for a used-car sales company. Your boss asks you to write a short report (a maximum of half a page) on which features (other than kilometers) could be modified or added to the cars in store that would increase the selling price. Your report should talk about the statistical evidence you have found in your analysis, but in terms that your boss can understand and put into practice.

2) **A not-so-traditional exercise.**
Notice the data in the second and third worksheets. The worksheet "**Housing Prices complete**" contains data on 250 houses that were sold in Barcelona a few years back. The data are **different characteristics of the house** (square meters, number of bedrooms, number of bathrooms, the neighborhood where the house is located, etc...) and the **price in euros** that the house sold for.

Your task is to **design and estimate** a regression model that will explain **the price of "Barcelonian" houses in terms of their main characteristics**. Then you are required to use this model to **predict** the price of other similar houses. Specifically, you will have to predict the prices of the 200 houses that you can find in the worksheet "**Housing Prices predict**". These are a random set of houses that were taken out of the original database. If your model is correct, it should predict reasonably well these *out-of-sample* prices.

Thus, the deliverable for this exercise is your **final regression model** (which you can simply write down in the Excel worksheet in the yellow cell), and the **predicted prices for the 200 houses** in "Housing Prices predict" in the column of yellow cells. Also, you need to give me

the **R2** of the model that you estimated for the 250 houses used in the analysis. Rename the excel file with your name and send it to javier.gomez@upf.edu.

I will then collect your 200 price estimates from the yellow cells and compare your **prediction** for each house with the **actual** price of that house. I will calculate the RMSE (root mean squared error) of your predictions to ascertain the accuracy of your model. There are some interesting learning points that come from this exercise and I will review them after you turn in the assignment. I ask you for two favors:

- Do NOT change the order of the rows in the 200 houses to be predicted (this will facilitate my analysis!).
- Do not worry about grading of this exercise, since you will get full credit for a reasonable submission regardless of the results. However, I need you to work on this exercise **individually** (so not in groups) and **independently**, so that your answers are not related to those of any of your colleagues. This will make the learning points more robust and interesting.

## Stata commands

For Stata, the command for linear regression is

**regress** *depvar* [*varlist*] …

where *depvar* is the name of the dependent variable and *varlist* is the list of independent variables to be used in the regression. There are lots of other options that go with the command (see the help files or the Stata manual or the notes from your brush-up course). I assume you already know how to input your data into Stata variables. Adding ", robust" or, simply, ",r" at the end will use heteroskedasticity-robust standard errors.

If you want to find the residuals from the regression, use the command:

**predict** *new_variable*, *residuals*

If you want to find the predicted values of the dependent variable, given your regression model, use the command:

**predict** *new_variable, xb*

which computes the predicted values of Y given X (the *xb* is not necessary, since by default Stata computes predicted variables if you do not specify anything else). If you want your prediction to be "out of sample", you need to have the values of the explanatory variables in a different datafile and then change the active file. Otherwise, Excel can easily do this for you, once you have the regression output.

If you want to have standard errors for the predicted values, use the commands

**predict** *new_variable*, *stdp*

**predict** *new_variable*, *stdf*

These two are slightly different, so read the Stata manual and see if you understand the difference (subtle and not so important in real life, but still it helps our understanding!)

Stata can do tests of linear hypotheses. The more common is, of course, testing that the coefficient of a specific variable is equal to 0:

**test** *variable1*

but you can test whether the coefficient on that variable is equal to some other value

**test** *variable1= 0.5*

or whether the coefficient on variable 1 is equal to the coefficient on variable 2

**test** *variable1=variable 2*

or you can test several hypotheses at the same time

**test** (*variable1=variable2*) (*variable3 = 0.5*) (*variable4*)

or you can test general linear combinations of parameters

**test** *2*(variable2-3*(variable3-variable4))=variable3+variable2*

If you want to test a nonlinear hypothesis (nonlinear in the parameters), then for example

**testnl** *2*_b[variable1]+3*_b[variable1]^2-_b[variable2]=0*

For heteroskedasticity, you can use:

**estat** hettest

which will perform the Breusch-Pagan test using fitted values of y as the possible explanatory variables for the heteroscedasticity (a simple version of White's test), or

**estat** hettest, rhs

which will use the original right hand side variables to test for heteroscedasticity due to each variable separate, or

**estat** hettest *variables*

which will perform the test only for the variables you specify. Also, you can use the imtest command with the "white" option, which will perform White's heteroskedasticity test:

**estat** imtest, white

**R commands**

For R, the command for linear regression is

reg_name <- **lm(**depvar~x1+x2+...+x4, data=df)

where *depvar* is the name of the dependent variable and *x1* to *xk* are the independent variables to be used in the regression. I assume you already know how to input your data into R datasets and into the dataframe df. Alternatively, but much more cumbersome:

reg_name <- **lm**(df$depvar~df$x1+df$x2+...+df$x4)

reg_name is the name you give to the model, and will allow you to call other commands afterwards (such as prediction errors, fitted values, hypotheses tests,…) and R will understand that it needs to take the output of that particular model.

**summary**(reg_name)

will give you the output of the regression model reg_name in a more standard format.

If you want to use robust (heteroskedasticity consistent) standard errors, you need to call the package *car* or the package *sandwich*. The default options (equivalent to Stata's robust standard error option) are:

**library(**car**)**
**rob.se1 <- hccm(**reg_name, type="hc0")

**library(**sandwich**)**
**rob.se2 <- vcovHC(**reg_name, type="HC0")

which then you can use as inputs of **coeftest**, for example:
**coeftest(**reg_name, rob.se1)

If you want to find the residuals from the regression, you can use the command:

**resid(**reg_name**)**

If you want to find the predicted values of the dependent variable you can use the command:

**fitted(**reg_name**)**

which computes the predicted values of Y given X. If you want your prediction to be "out of sample", Excel can easily do this for you, once you have the regression output or you can use the command **predict**:

**predict(**reg_name, cvalues)

where cvalues is a dataframe which contains the values of the regressors for the out-of-sample individuals.

If you want to have standard errors for the predicted values, you can include the options:

**predict(**reg_name, cvalues, interval="confidence") -> for standard errors of the predicted mean

**predict(**reg_name, cvalues, interval="prediction") -> for standard errors of the predicted value for one individual

These two are different, so you can read the R manual and see if you understand the difference (subtle and not so important in real life, but still it helps our understanding!)

R can do tests of linear hypotheses in different ways (the command **coeftest** does this for the basic tets of all regression coefficients).
The more flexible command is **linearHypothesis** (from the **car** package), which gives you plenty of options. You need to specify a vector **myH0** which contains your hypotheses:
**linearHypothesis(**reg_name, myH0**)**

The simplest test is, of course, testing that the coefficient of a specific variable is equal to 0:
**myH0 <- c(**"X1"**)**
where X1 is the name of the variable whose coefficient you want to test, but you can test whether the coefficient on that variable is equal to some other value
**myH0 <- c(**"X1=0.5"**)**
or whether the coefficient on variable 1 is equal to the coefficient on variable 2
**myH0 <- c(**"X1=X2"**)**
or you can test several hypotheses at the same time
**myH0 <- c(**"X1", "X2=0.5", "X3=X4"**)**
or you can test general linear combinations of parameters
**myH0 <- c(**"X1-3*X2=X4"**)**

For heteroskedasticity tests, you can use:
**library(**car**)**
**bptest(**reg_name)
which will perform the Breusch-Pagan test using all regressors, or
**bptest(**reg_name, ~regressors)
which will perform the general form of White's test, or
**bptest(**reg_name, ~fitted(reg_name)+I(fitted(reg_name)^2))
which will perform the simple form of White's test (also a form of a BP test).