Spring 2025 ENVECON/IAS 118 - Introductory Applied Econometrics Problem Set 1

Due on Gradescope, Midnight February 7

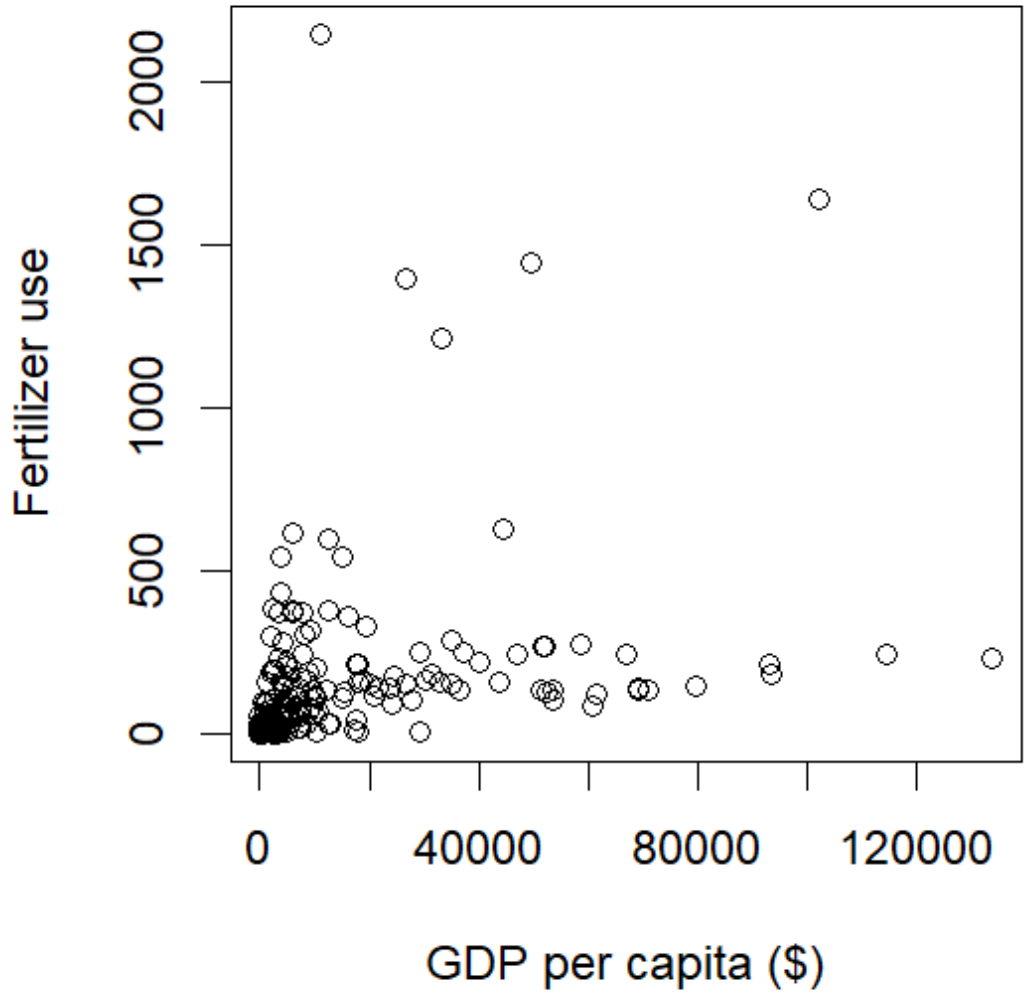
Submission Instructions

1. Complete the problem set using the provided Jupyter Notebook template (on datahub). Make sure to save your work as you go!
2. When you are done, export your submission as a PDF. Make sure that all of your code and your output (figures, calculations, etc) are included in your final submission to receive full credit (run all of your cells!).
3. To export as a PDF, go to the file dropdown menu and select the ”Save and export notebook as” dropdown menu. In this menu make sure to select ”PDF”, "Webpdf" or "PDF via Chrome" (if that option appears instead).
4. Upload your completed submission to Gradescope: https://www.gradescope.com/courses/927499.

Note: You can also complete this problem set using R studio directly. If so, use R Markdown to create a PDF for your submission. Your submission must include all code and output in order to receive full credit.

As you see in Figure 1, there appears to be some association between GDP per capita and fertilizer use in Agriculture.

Figure 1.



The data are from the World Bank development indicators (http://data.worldbank.org/data-catalog/world-development-indicators) for 2021, except for Eritrea and South Sudan. Fertilizer use is measured as the consumption of fertilizer on arable land (land used for cultivation or for pasture). It is measured in tons per hectare of arable land. GDP per capita is gross domestic product divided by midyear population, and is also measured in US $. Figure 1 above includes all 189 countries from the original dataset, for which we have GDP per capita and fertilizer use.

We only include some (randomly) selected countries for the assignment. The values for selected countries can be found in the csv files "countriesA_fertilizer.csv" and "countriesB_fertilizer.csv". We will use both datasets.

In [1]:
```
#(optional formatting) get rid of scientific display of numbers
options(scipen = 100, digits = 4)
```

Exercise 1. Relationship between GDP per capita and Fertilizer Use

We will estimate a simple linear relationship between Fertilizer use and per capita GDP on a subset of 5 countries.

(a) Use R to create a scatter plot of these observations.

a-Step 1: Load the .csv file called countriesA_fertilizer.csv. (Hint: the `read.csv()` command will likely be helpful.)

a-Step 2: Look at the data. This dataset only has 5 rows so you can just call the entire dataset. In general you want to use the `head()` command so that R does not print the entire dataset which will take way too many pages.

a-Step 3: Rename the variables to "countryname", "gdp_pc", and "fert_use". (Hint: the `colnames()` command may be useful. Also remember that to select multiple values (such as mulitple variable names, you can use R's vector notation `c()`. For example: `c("a", "b", "c")`.

a-Step 4: Create a scatterplot of the data. Make sure to (1) label the axes and their units, and (2) title your graph. (Hint: the `plot()` command will likely come in handy. Use `help(plot)` or `?plot` to view the documentation for the function and how to include labels.)

b) Estimate the linear relationship between GDP per capita and Fertilizer consumption ("F") by OLS, showing all intermediate calculations.

$$\widehat{F} = \hat{\beta}_0 + \hat{\beta}_1 GDPcap$$

For this exercise, **DO NOT** use the built-in R commands like `cov()` or `lm()`. Use basic mathematical commands (`+`, `-`, `*`, `\`, `sum()`, `^`) to produce all the values and show all the steps.

b-Step 1: Create new data objects called **mean_gdp_pc** and **mean_fert_use** equal to the mean of **gdp_pc** and **fert_use**.

b-Step 2: Calculate the covariance (only using the mathematical operations specified above) between gdp_pc and fert_use: $cov(gdp_{pc}, fert_{use})$.

- Do this by first creating two new columns of residuals: **resgdp**, a column that subtracts the **mean_gdp_pc** from **gdp_pc** and **resft** that subtracts the **mean_fert_use** from **fert_use**.
- Next create a column **resgdpft** which is equal to **resft** multiplied by **resgdp**.
- Finally, generate a value named `covarA` which is equal to the sum of **resgdpft** divided by n-1.
- Make sure to call `covarA` at the end so we can see it printed in the output.

Hint: To add new columns to your dataset, you can either use mutate as in Small Assignment 1 or you can use the following syntax: `dataset_name$new_var_name <- formula for the new variable`. Another option is to use cbind (as explained here: https://www.statology.org/r-add-a-column-to-dataframe/)

b-Step 3: Calculate the variance.

- First generate a column **sqresgdp** equal to the square of **resgdp**.
- Generate a value named `varA` which is equal to the sum of **sqresgdp** divided by n-1.
- Make sure to call `varA` at the end so we can see it printed in the output.

b-Step 4: Using the quantities generated above, generate and print `beta_1hat` and `beta_0hat`, your estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$.

c) Interpret the value of the estimated parameters $\hat{\beta}_0$ and $\hat{\beta}_1$

d) In your data frame, compute the fitted value and the residual (the difference between the actual and fitted value) for each observation. Use only basic mathematical commands (`+` , `-` , `*` , `\` , `sum()` , `^`) to do this. Create a new column named "fitted" and another new column called "residuals". Call the head() of your dataset so we can see these new columns. Verify that the residuals sum to 0 (approximately).

e) Now use the `lm()` command to run this regression automatically rather than manually as you did above and save the output as "reg1".

Check that your estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ that you calculated manually above match the estimates using `lm()`. Call the `summary()` of reg1 so we can see the output.

f) According to the estimated relationship, what is the predicted $\widehat{F}$ for a country with a GDP per capita of $13,000?

g) How much of the variation in fertilizer use for the 5 selected countries is explained by their GDP per capita?

Calculate the R^2 by calculating the sum of squared model residuals and the sum of squared total (variation of the dependent variable). Use only basic mathematical commands (`+` , `-` , `*` , `\` , `sum()` , `^`) to do this. Then calculate R^2 and make sure to call the value so we can see it printed out.

h) Repeat exercises (a) and (e) for the additional set of countries whose data is available in the file countriesB_fertilizer.csv.

i) How do your estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ change between the two sets of countries? Discuss and briefly explain this variation in 2-3 sentences.

Exercise 2. Functional Forms

Suppose you estimate an alternative specification given below for sample B countries, where log() refers to the natural log, ln().

$$\widehat{F} = \hat{\alpha}_0 + \hat{\alpha}_1 log(GDPcap)$$

Let

$$\hat{\alpha}_0 = -315.6$$

and

$$\hat{\alpha}_1 = 52.3.$$

a. Please give at least one reason discussed in lecture/section why we might log GDPcap in our regression model.

b. What is the predicted quantity of fertilizer used for a country with a GDP per capita of 10,00 US$?

c. How would you interpret the marginal effect of GDPcap on Fertilizer use based on your estimated model ($\hat{\alpha}_1 = 52.3$), being careful to take the functional form of your model into account?

Hint: Look at Section 1