Data Modelling & Analytics
Individual Assignment 3

INSTRUCTIONS

- This is an individual assignment.
- Submit your answer digitally as **a single pdf file** (file extension: pdf) through Canvas.

Deadline: Monday, May 30 at 23:59.

- Late submissions are not allowed.

Warning: The detection of any form of plagiarism in your work means the assignment will be graded with ZERO points.

Load and prepare the data

Load the data "intdef" from package "wooldridge" and the data "Gasoline" from package "plm".

As for the data "intdef":

- Please see slide 5 of video 5.1 and the following link for the description of variables: https://rdrr.io/cran/wooldridge/man/intdef.html .
- Create a new data set that is the same as the data "intdef" but with the name "intdefxxx". "xxx" should be the last 3 digits of your student number. For example, if one student's student number is 123456, then the name of the new data set is "intdef456".
- Rename the variables year, i3, inf, and def in the data "intdefxxx" to be yearxxx, i3xxx, infxxx, and defxxx. For example, if one student's student number is 123456, then the new names of the variables are year456, i3456, inf456, and def456.
- For the rest of individual assignment 3, please use the data "intdefxxx".

As for the data "Gasoline":

- Please see the following link for the description of variables: https://rdrr.io/cran/plm/man/Gasoline.html .
- Create a new data set that is the same as the data "Gasoline" but with the name "Gasolinexxx". "xxx" should be the last 3 digits of your student number. For example, if one student's student number is 123456, then the name of the new data set is "Gasoline456".
- Rename variables country, year, lgaspcar, lincomep, lrpmg, and lcarpcap in the data "Gasolinexxx" to be countryxxx, yearxxx, lgaspcarxxx, lincomepxxx, lrpmgxxx, and lcarpcapxxx. For example, if one student's student number is 123456, then the new names of the variables are country456, year456, lgaspcar456, lincomep456, lrpmg456 and lcarpcap456.
- For the rest of individual assignment 3, please use the data "Gasolinexxx".

Please note that in the rest of this instruction, "xxx" refers to the **last 3 digits** of your student number.

The principle is that the names of all the variables and data sets that are used by you to finish parts A and B of this individual assignment 3 should include "xxx". The sum of all points in parts A and B is 25. Assignment reviewers will first do grading according to the model answers of parts A and B, and then deduct 5 points if this principle is violated.

Part A: Time series

For part A, please use the data "intdefxxx".

1. [0.8 points] Define variable defxxx from the data "intdefxxx" as a ts object and plot it.

2. [1 point] Define "intdefxxx" as a zoo object containing all data. Make a time series plot of variable defxxx. Compare this plot with the one that you make for question 1. Are the two plots the same?

3. [2.5 points] Use the command "lm" to fit a finite distributed lag (FDL) model of order 3:
   - Dependent variable: i3xxx
   - Independent variables: infxxx, defxxx, defxxx lagged by one time unit, defxxx lagged by two time units, defxxx lagged by three time units

   Compare this FDL model of order 3 with the static time series model in slide 5 of video 5.1. Which model should you choose, this FDL model of order 3 or the model in slide 5? Please provide an explanation about how you make the decision.

4. [2 points] Use the command "dynlm" to fit a FDL model of order 3 with the same dependent variable and independent variables as those in question 3:
   - Dependent variable: i3xxx
   - Independent variables: infxxx, defxxx, defxxx lagged by one time unit, defxxx lagged by two time units, defxxx lagged by three time units

   Use the command "stargazer" to make a table of regression results in questions 3 and 4:
   - column (1) shows the result that you get in question 3 using the command "lm"
   - column (2) shows the result that you get in this question using the command "dynlm"

   Does command "dynlm" give you the same result as command "lm"?

5. [1.5 points] Test whether you should add a time trend in the above FDL model of order 3. Specifically, you compare the model in question 4 with the following FDL model:
   - dependent variable: i3xxx
   - independent variables: infxxx, defxxx, defxxx lagged by one time unit, defxxx lagged by two time units, defxxx lagged by three time units, and time trend

   Which model should you choose, the FDL model with a time trend or without a time trend? Please provide an explanation about how you make the decision.

6. [2 points] Based on your chosen model in question 5, calculate the estimated value of long-run propensity (LRP) of variable defxxx. Test whether this LRP is significant. Interpret LRP.

Part B: Panel data

For part B, please use the data "Gasolinexxx".

1. [0.2 points] Which variables are the entity index and time index of the panel data "Gasolinexxx"?

2. [1 point] Create a new variable (a new column) called "m_lincomepxxx" in the data "Gasolinexxx" such that, for every entity, the value of the variable "m_lincomepxxx" is the mean of lincomepxxx across different years.

3. [1 point] Define the data "Gasolinexxx" as a panel data frame in R. What are panel dimensions? What are the meanings of the numbers n, T, and N in RStudio console output? What are the time-invariant and individual-invariant variables of this panel?

4. [2.5 points] Make the following two plots:
   - Plot 1: A plot of dependent variable lgaspcarxxx and yearxxx for every entity
     Hints: The format of the plot should be similar to the example plot presented in the videos. In the plot, there should be a line for every entity. Make sure that the labels of your plot are clear to see.
   - Plot 2: A plot for fixed effects: Heterogeneity across entities
     Hints: The dependent variable is still lgaspcarxxx. The format of the plot should be similar to the example plot presented in the videos. In the plot, there should be black points and a red line. Make sure that the labels of your plot are clear to see.

   What do you observe from the two plots? Please describe the two plots. Based on the two plots, do you think whether the individual fixed effects should be taken into consideration or not? Please explain why you think the individual fixed effects should or shouldn't be taken into consideration.

5. [2.5 points] Use the command "lm" to fit a least squares dummy variable (LSDV) model that considers individual fixed effects:
   - Dependent variable: lgaspcarxxx
   - Independent variables: lincomepxxx, lrpmgxxx, lcarpcapxxx, etc.

   Use the command "plm" to estimate a FE estimator (or within estimator) that considers individual fixed effects:
   - Dependent variable: lgaspcarxxx
   - Independent variables: lincomepxxx, lrpmgxxx and lcarpcapxxx

   Use the command "stargazer" to make a table of the results in this question:
   - column (1) shows the result of the LSDV model
   - column (2) shows the result of the FE estimator
   - only include variables lincomepxxx, lrpmgxxx, and lcarpcapxxx in the table

Compare the result of the LSDV model and the result of the FE estimator. Do you get the same estimated coefficients and standard errors of variables lincomepxxx, lrpmgxxx and lcarpcapxxx?

6. [0.5 points] Instead of using variable lincomepxxx as an independent variable, Lucy wants to use m_lincomepxxx as an independent variable. Can she get an estimated coefficient on the variable m_lincomepxxx if she uses a FE model, yes or no? Please provide an explanation for your answer.

7. [1.5 points] Should you add time fixed effects to the FE model in question 5? In other words, should you choose a FE estimator with both individual and time fixed effects or only with individual fixed effects? Please provide an explanation about how you make the decision.

8. [4 points] In the following analysis, the dependent variable is still lgaspcarxxx. The independent variables are based on the model chosen by you in question 7:
   - If in question 7 you choose a model that doesn't include time fixed effects, then in the following analysis, your independent variables are lincomepxxx, lrpmgxxx, and lcarpcapxxx.
   - If in question 7 you choose a model that includes time fixed effects, then in the following analysis, please take the time fixed effects into consideration by including year dummies as your independent variables in your regressions. So your independent variables are lincomepxxx, lrpmgxxx, lcarpcapxxx, and year dummies.

   Estimate a pooled OLS model, a FE model, and a RE model using the above dependent variable and independent variables. Use the command "stargazer" to make a table of the results:
   - column (1) shows the result of pooled OLS
   - column (2) shows the result of the FE model
   - column (3) shows the result of the RE model
   - only include variables lincomepxxx, lrpmgxxx and lcarpcapxxx in the table

   Which model should you choose among pooled OLS, FE model, and RE model? Please provide an explanation about how you make the decision. Based on your final chosen model, is the coefficient on lrpmgxxx significant or not?

9. [2 points] Test whether there is considerable serial correlation in your chosen model of question 8. Based on the test, should you use the standard errors that you get in question 8 or the robust standard errors? Please provide an explanation about how you make the decision. Based on the decision, will you change your conclusion about whether lrpmgxxx is significant or not?
   If you decide to use robust standard errors, please calculate the robust standard errors and use the command "stargazer" to make a table of your results:
   - column (1) shows the result of your chosen model with standard errors in question 8
   - column (2) shows the result of your chosen model with robust standard errors

Format of your answers

For every question, after running your R codes and seeing the output in the console in RStudio, copy and paste everything from your **console output** to the **"grey" part** in the word file "**Individual assignment 3_Format template.docx**". Select everything (including the R codes and the regression results, etc.) and change the font size to 10 (this should be automatically done if you choose "Keep Text Only" when you paste). **Plots** should be copied and pasted in the "grey" part following the R codes that are used to create them. **Tables** created by the command "stargazer" should be copied and pasted below the "grey" part and should look the same as they are shown in the RStudio console. In sum, the "grey" part should contain the plots and everything from your RStudio console output, except the tables created by the command "stargazer".

You can put your mouse cursor in the "grey" part and press the key "Enter"/"Backspace" to extend/reduce the length of the "grey" part. You should also write **the rest of your answer** to the question below the "grey" part, e.g., your explanation about why you choose a specific model, your answer to the yes or no question, etc.

After you have everything in the correct format in the word file "Individual assignment 3_Format template.docx", save it as a pdf file with your **student number** (e.g. 123456AB-assignment3.pdf) and submit it through the Canvas.

I have created a word file "**Examples of the correct format.docx**" with two example questions and the corresponding model answers to show you the correct format. Assignment reviewers will first do grading according to the model answers of parts A and B, and then deduct 5 points if your final submitted pdf file has a wrong format (e.g., R codes or output in RStudio console are not in the "grey" part, the tables created by the command "stargazer" don't look the same as they are shown in the RStudio console, forgetting to copy and paste ">" symbol for every line of R codes, naming your pdf file without your student number, etc.).

Note: There is a limit on the number of lines the console output can save. Thus, you should copy all the output in time, so that you do not lose work.