1 General Information For The Extra Task

Task Worth	An extra 10 points to the total (roughly an additional 10% of the maximum grade)	The extra task is optional - the maximum total points for this course could be: 110 = Midterm I + Midterm II + Exam + Extra Task
Deadline	~~2018-12-13~~ **2018-12-21** (extended)	**Late submissions not accepted.**
Team/group size	**up to 4 people**	Individually form the groups amongst yourselves
Task assignment	No duplicates! Any one task can be done by one team only - **multiple teams cannot do the same task**.	Prepare a list of teams and their selected tasks.
Software	R and Python	Both programming/scripting languages must be used to showcase what can be done (or indicate what functionality is lacking in one language, but is awailable in the other)
Citation	Provide references and sources as either book titles and authors, or in case of open-access sources - website links.	This includes links for theory, software libraries, empirical datasets or other code examples used.
•	•	•
Quality of the work	If the task is done exceptionally well, more than 10 points may be awarded.	This includes multiple examples with explanations, comparison (pros & cons of the functionality for the discussed methods in R and Python).
•	•	•

2 File requirements

File	Fyle Types	Info
Technical Document	.ipynb file + the **compiled** .html **or** .pdf file. **Note:** • If possible, try to limit the methodology to cover cross-sectional data only; • If there are multiple estimation methods, select 1-2, which you understand and can carry out in R and Python. Then only mention the existence of the remaining estimation methods and their pros/cons.	Should contain the main theoretical background for the task. **Proofs and derivations are not needed** **The beginning of the document should indicate what each member of the team contributed in this task (a couple of sentences from each member).** This document should provide the background on the presented methods along with some formulas on any model specifications, estimations, hypothesis testing, or similar algorithm theoretical framework used in the task. If the general formulas are complex, you are free to present a simplified formula for a particular case (e.g. formulas for a model with one independent variable, normally distributed residuals, etc.). It should be clear from reading the document, that the presented theory (and/or its simplified version) was understood and that the presented methodology was applied in the remaining files;
R code example document	.ipynb **or** .Rmd file + the **compiled** .html **or** .pdf file.	Format the file in a readable way: the examples should be provided using blocks of formatted text, mathematical formulas and explanations. If only the code is provided with some *#comments* but no formatted explanations, formulas or examples, then **this file is not regarded as submitted** . You should be able to run the code (much like in the example files for the lectures) from within the .ipynb, or .Rmd files.
Python code example document	.ipynb file + the **compiled** .html **or** .pdf file.	Format the file in a readable way: the examples should be provided using blocks of formatted text, mathematical formulas and explanations. If only the code is provided with some *#comments* but no formatted explanations, formulas or examples, then **this file is not regarded as submitted** The provided examples should be the same as in R. If using an empirical dataset - the same dataset should be used in both R and Python. If simulating data - simulate with the same mathematical formulas and parameter/distribution assumptions for both R and Python.
Empirical data	.txt, .csv file, or a link to download the data	If you are using an empirical dataset, sent it separately, or load it from a url in the code documents. The data should be freely available (open-access).

3 Some Considerations When Preparing the Relevant Files

Try to answer the address the following points when preparing the relevant files:

- **Technical Document:**
  - ☐ Briefly describe the topic that is used:
    - ☐ For what reason would you would want to use the methods in this topic?
    - ☐ What kind of data is required?
    - ☐ What questions do these methods, or the overall topic, help answer?
  - ☐ Briefly give the mathematical background for the methodology:
    - ☐ What are the assumptions for the variables and the data in the methodology?
    - ☐ What are the main formulas used and what are they used for (model specification, parameter estimation, dataset stucture, etc.)?
    - ☐ Can you provide the asymptotic properties - the distribution, mean and variance of the parameters, estimation methods, etc.?
    - ☐ Can you provide a minimalistic theoretical example (e.g. how would the parameter estimation formulas look like for a theoretical dataset with one explanatory variable, or one characteristic group, etc.)?
    - ☐ Are there any tests, or visuals (plots, histograms, etc.) that are helpful in this methodology?
  - ☐ Briefly compare with some other methods, models, or simply overall performance of the method:
    - ☐ What are the pros of this methodology (may be similar to the reason for using these methods)?
    - ☐ What are the drawbacks of this methodology (if any; may include difficulty to estimate, lack of precision, non-interpretable coefficient values, etc.) ?
    - ☐ How does it compare to some similar methods? (Note: no need to go into depth for the other methods - especially if these methods are outside of the ones discussed during the lectures - simply give their names and link to relevant sources).
  - ☐ **(Optionally)** Some additional points to consider:
    - ☐ Have you found any criticisms of this methodology (maybe it was important in the past, but is heavily criticized nowadays)?
    - ☐ Have you found any data examples (e.g. some specific economic industries, medicine, biology, etc.), which can use this methodology? What does this methodology help answer/identify in those examples?
- **R and Python files:**
  - ☐ Dataset (ideally both, but **doing only one is enough**) to highlight the methodology:
    - ☐ Can you simulate data for an example? If no, then provide a reason. Otherwise, provide the assumptions/models/formulas for the data simulation in a mathematical form and the code, which simulates the data.
    - ☐ If you cannot simulate the data - use an empirical example dataset.
  - ☐ **(Not required)** Can you simulate an example, when the methodology doesn't work?
  - ☐ Carry out the methodology, outlined in the technical document (tests, plots, model specification, estimation, etc.):
    - ☐ Use the built-in functions in R and Python. If some functionality is not available - indicate what features are missing.
    - ☐ **(Not required)** Manually carry out (some parts of) the methodology to show how some calculations can be done. Compare them with the built-in functions. If the results differ - provide an explanation for why this is the case.

Additional Notes:

- The files should be inclusive - i.e. you should be able to run the code from the beginning to the end of the raw file to either simulate, or read-in, the data and carry out the methods presented.
- The text should be **cohesive** - when using multiple references - formulas and variables should be unified across different sources.
- **You do not need to provide proofs or derivations for the formulas** (though they may sometimes be helpful), as long as you provide the sources for the formulas.
- **You do not need to have an answer every single point** (again, do not treat it as a questionnaire). For example, if parameter estimation does not have a concrete expression - as is sometimes the case - writing down the formula and indicating that optimization is done via numerical optimization (indicating the optimization algorithm as well) is enough (as long as you can back this claim with a source/reference).
- **You do not need to cover every single estimation method**. Usually there are 1 or 2 estimation methods which are considered the *best* in terms of their accuracy/complexity/calculation time. In such cases, give a list of the methods, which **could** be used, along with their pros/cons. Then, select one (or two) estimation methods and give their mathematical formulas, etc. as mentioned in the requirements. **Make sure that your selected methods can be carried out in R and Python!**

Things to avoid:

- **Useless code** - only estimate the models, create variables, and so on if you want to present their results, compare them with other models/data, draw some other kind of conclusions, or highlight some method properties;
- **Incomplete code**, errors in the code - if the code does not work it should not be included.
- **Code, which you do not understand**. This also results in the previously mentioned errors in the code. Complex data preparation and plotting code also frequently results in unresolved errors.
- **Broken files**. If neither the raw, nor the compiled files are usable - the files are not regarded as submitted.

4 Tasks

	Topic	Theory	Examples
(1)	Quantile Regression	Wiki, [1], [2], [3], [4], [5]	[1], [2], [3]
(2)	Multilevel model (HLM)	Wiki, [1], [2], [3], [4], [5], (similar into: [6], [7], [8], [9],)	[1], [2], [3], [4] (more advanced: [5], [6], [7])
(3)	Principal Component Regression	Wiki, [1], [2](p.230), [3](p.79(98)), [4]	[1], [2], [3], [4], [5], [6] (sim. ex.: [P1], [P2], [P3])
(4)	Local Regression (LO(W)ESS)	Wiki, [1], [2](p.280), [3](p.194(213)), [4] (note the math in each step)	[1], [2], [3], [4], [5]
(5)	k-means Clustering	Wiki, [1], [2](p.386),[3](p.460(479)),	[1], [2], [3], [4]
(6)	Hierarchical clustering	Wiki, [1], [2](p.394), [3](p.520(539)),	[1], [2], [3], [4], [5], [6], [7], [8]
(7)	Regression Trees	Wiki, [1], [2](p.304), [3](p.307(326))	[1], [2], [3], [4], [Video ex. in R],
(8)(?)	LASSO[1]	Wiki, [1](p.219), [2](p.68(87))	[1](p.255)

Note: each task/topic can only be done by one group.

[1] **Optionally** the LASSO topic can also cover Lasso + Ridge regression, as Rigde regression is in the same chapters of the referenced books. This is not necessary, but can make it easier to present the background of the model and compare results, etc. Again, this is not a requirement for this topic.