Assessment of Master Thesis – Opponent

Study programme: **Quantitative Methods in Economics**
Field of study: **Quantitative Economic Analysis**
Academic year: **2018/2019**
Master Thesis Topic: **Model building in regression analysis**
Author’s name: **Bc. Kristina Lobanova**
Ac. Consultant’s Name: **Mgr. Milan Bašta, Ph.D.**
Opponent: **Ing. Karel Helman, Ph.D.**

	Criterion	Mark (1–4)
1.	Comprehensibility of the Master´s Thesis topic and objectives	**1**
2.	Fulfilment of defined objectives	**1**
3.	Logical structure and cohesion of each parts	**1**
4.	Extent and relevance of description of the current level of knowledge	**2**
5.	Adequacy of methods in respect of the topic (selections of the methods and their application)	**1**
6.	Extent, quality and precism of description of the thesis´s results	**1**
7.	Relevance and correctness of discussion of the thesis´s results	**1**
8.	Correctness and relevance of information sources	**1**
9.	Grammar, stylistic style, terminology and overall formal and grahic level of the Master´s thesis	**1**

Comments and Questions:

I enjoyed reading the present thesis very much. It illustrates many interesting features an analyst should bear in mind when employing the regression analysis (concretely multiple linear regresson model) in any social or scientific field of interest. Undoubtedly, the author showed an ability to perform a complex statistical analysis on her own as well as profound competence in interpreting achieved results, both on a level reaching (according to my experience) above the usual level of comparable master theses.

There are many interesting results (outputs) presented in the practical part of the thesis in form of tables and figures, when wide variety of theoretical concepts and tools is applied during construction of an interesting multiple regression model. The author conducted her analysis using a dataset of 183 countries, the dependent (response) variable being the life expectancy.

I would like also to explicitly highlight the way the author worked with sources, effectively combining many interesting sources, some of them very up-to-date.

I missed a bit a list of used symbols. I suggest avoiding too strong statements like „must“ or „necessary“ when dealing with some steps in a statistical analysis as well as using too „subjective“ terms without detailed explanation what they mean (e.g. „satisfactory“ on page 1, „suitable“ on page 6 etc.). Sometimes, the text of the thesis is not completely clear, provoking question like „in what sense?“, „why?“ etc. Adjusted R-squared has no factual interpretation (as is given on page 58).

Mine other comments are to be understood as suggestions or stimulations for future work rather than a criticism of the thesis as some of them are very intricate ones (far above usual level of master theses).

Firstly, throughout the whole thesis, an important issue is „estimation of population parameters“. In the conclusion (page 84) it is said that „Based on a sample drawn randomly from a population, the OLS approach is employed…“. The concern of mine, however, during the whole thesis, was: using a dataset of 183 countries, what would be the population? Or in other words, from what group of elements (countries) was the analyzed dataset drawn as a random sample? In my opinion, even though there are many conlusions about a population made in the thesis (in fact all the statistical hypothesis tests and confidence intervals), the discussion and advocacy of what group of elements is to be understood as the population, is completely missing. Another aspect of this issue is that if the explanatory variables are considered to be non-random (as it seems to be so at some spots in the thesis), all the statistical inference (tests and intervals) are (strictly speaking) not about any population but about a data-generating process.

Secondly, the main purpose/goal of the multiple regression model constructed in the thesis is unclear. Sometimes it seems that it is intended to serve for explanatory/causal conclusions (e.g. page 62), sometimes it seems to aim on predictions (mainly chapter 10). For different aspects of explanatory/causal and predictive regression modelling see e.g. „To Explain or to Predict“ (Shmueli, 2011).

Lastly, the relationships among the explanatory variables. I like chapter 6 very much, mainly its section 6.2 where the expected individual impact (individual relationships) of each explanatory variable on the response variable (life expectancy) is studied. In the multiple regression model (model with more explanatory variables), however, the simultaneous (i.e. very different) impact of all explanatory on the response variable is studied, the partial regression coefficients showing partial effect of each explanatory variable only. I think that the individual and partial effect would be the same only in case of complete independence among the explanatory variables, otherwise those effects may differ a lot. Some discussion on the relationships among the explanatory variables could thus be also useful/interesting.

For a discussion during defending of the thesis I suggest some of these questions:

1. It is clear enough what does symbol „y_i“ signify in the empirical regression model (page 7). Nevertheless, in the theoretical regression model (page 4), it should be something quite different… What does „y_i“ signify in the theoretical regression model on page 4? And what is „i“ and „n“ there?
2. Does really the “FGLS estimation measure the marginal impact of each X_j on y” as is said on page 17, even though the original data were transformed?
3. At page 9 in A6 assumption (normal distribution of the error term) it is said that “…this is only concern when the sample size is very small. When the sample size is sufficiently large, the CLT ensures that the distribution of the unobservables will be approximately normal”. Then, on page 59, many statistical tests are carried out for a model constructed using data about 183 countries. Does this mean that 183 is a very small sample size, so that CLT does not work?
4. In case of changing the dataset (e.g. by removing outliers from the analysis) does the population (to which all the generalizing conclusions are drawn) change as well?
5. How exactly was the “Income level” incorporated in the constructed regression models? Was it somehow just one dummy variable as could be understood from page 86?
6. On page 56 you state that “…since this value does not dramatically exceed the threshold…”. The value being 10.79, the threshold being 10. What would have to be the value for you to conclude that it (dramatically) exceeds the threshold (i.e. 10)?
7. Is Figure 11 on page 60 ok? It seems like the fitted values of life expectancy are the highest for low-income countries…
8. What would be your conclusion if there was very strong individual relationship between an explanatory variable and the response variable but p-value (of the tests commented on page 65) would be higher than 0.1?

My proposition of the final mark below assumes at least basic ability to react on these questions.

Conclusion: The Master Thesis is recommended for the defence.

Suggested Grade: **1**

Date: 06/08/2019

Ing. Karel Helman, Ph.D.
Opponent