Assessment of Master Thesis – Academic Consultant

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**
10.	Student's independence during the process of composing the Master thesis	**1**

Comments and Questions:

The goal of the thesis is to outline the main steps used in building a regression model and to practically implement the steps using real-life data.

All the goals which have been set up in the introduction part of the thesis have also been met. The notion of the classical linear regression model and its assumptions have been introduced. The method of least squares has been outlined together with the properties of the least-squares estimators. Statistical inference has been discussed including the F test for the overall significance of regression, individual t tests, univariate confidence intervals as well as joint confidence regions. The text continues with a nice assessment of the assumptions of a correctly specified regression function, homoskedasticity and normality of errors. Further, it provides a nice description of leverage points, regression outliers and influential observations as well as methods of their detection. Moreover, the method of weighted least squares (used also in the practical part) is introduced. I welcome the use of several books on regression analysis which resulted in a nice review of regression analysis in the theoretical part of the thesis.

I appreciate very nice real-life data used in the practical part of the thesis. All the variables used in the regression analysis are very nicely and thoroughly described. Expert knowledge is used to explain what the relationship between life expectancy and the explanatory variables may/should look like. All the results are illustrated with the use of very nice tables, graphs and figures. I appreciate that the issue of missing data has also been addressed in the analysis. I also appreciate the implementation of the weighted least squares in the analysis (even though I would say – based on Figure 11 – that the errors in the original model are practically homoskedastic and that the use of weighted least squares is not necessary). Several further techniques such as bootstrap, leverage points, externally studentized residuals etc. have been applied in the analysis.

The thesis is an original work (iThenticate has found some similarity with other documents, but the similarity is in "general sentences" only or with resources which are cited by the student – in other words, there is no doubt that the thesis is an original work).

My rating of the thesis is the highest possible (excellent). I have also some remarks. These are rather meant as recommendations or ideas for any similar work which the student may pursue in the future.

- I assume that the purpose of the regression model built in the practical part of the thesis is not only to find associations between life expectancy and the explanatory variables but to reveal causal factors that influence life expectancy. Causal modeling is always a difficult task and a reader can almost always be reserved about some results. I myself would suggest the inclusion of variables related to the type and level of education people receive in different countries – these variables are missing in the current model. On the other hand, I am a little bit reserved about the inclusion of adult mortality into the regression model since I do not understand the role of this variable if the purpose of the model is to find causal links. Moreover, in my opinion it could have been useful to run at least a brief literature review on factors that make people in different countries live longer or shorter lives. This review could have suggested what variables should be used in the model and in what form they could enter the model.

- What I am a little bit missing in the practical part of the analysis is practical interpretation of the results in terms of causality effects, the discussion whether any interactions could be present etc.? In some sense the practical analysis is more 'technical' but the reader may ask at several instances why this particular part of the analysis is performed (e.g. what is the practical reason for the construction of joint confidence region in Figure 14 etc.).

- As mentioned above, I appreciate the use of weighted least squares in the thesis. Its use in the detection of outliers, of influential observations etc. is – in my opinion – an advanced topic. Consequently, I would be a little bit reserved about some results obtained in the thesis in these parts and recommend further exploration of this topic in any potential student's future research.

- I would be careful in the interpretation of the effect of GDP and the signs of the estimated parameters on page 57. Rather than to interpret the sign of both terms (linear and quadratic separately) it could be more useful to address the joint effect of the two terms (linear and quadratic).

- There is perhaps no need to devote such a big space (2 pages) to the issue of orthogonal/raw polynomials and to calculate variance inflation factors to illustrate the difference between orthogonal and raw polynomials.

- There are some minor typos in notation (e.g. below equation 1.33 a different notation is used for the covariance matrix of parameter estimators; sometimes X is in bold sometimes not, see e.g. text around Equation 3.8). Some notation is not explained (e.g. what is R_j^2 in 1.31). There are some minor mistakes in Equations (e.g. in the in-line equation below Equation 3.1).

I would like to ask the student to answer the following questions during the thesis defence:

1. Is there any situation where the use of externally studentized residuals will fail to detect the presence of regression outliers (as defined in the thesis)? How shall we proceed in such situations in the detection of regression outliers?

2. Could you please comment on the effect GDP on life expectancy in the model of Equation 7.3.

3. You state on page 41 that the total number of regression models fitted during best subset regression is 2k, where k is the total number of predictors. This is presumably a typo. What is the correct number of models and why? Similarly, you state on page 41 that the criterion which can be used in best subset regression is R^2. This is presumably a typo again. What is the correct criterion and why?

4. Could you be a little bit more specific about what the null hypothesis in Equation 3.2 exactly states? Could you please explain why fitted values cannot enter into equation 3.4 linearly?

5. Is it possible to have non-normal errors while 'usual' inferential procedures remain valid at the same time?

The goals of the master thesis which the student has set up have all been met. I suggest grade 1 (= excellent) for the thesis.

Conclusion: The Master Thesis is recommended for the defence.

Suggested Grade: **1**

Date: 11/08/2019

Mgr. Milan Bašta, Ph.D.
Academic Consultant