UNIVERSITÄT BASEL   February 16, 2022
Faculty of Business and Economics
Prof. Christian Kleiber

Guidelines for the Preparation of an MSc Thesis
at the Chair of Econometrics and Statistics

Preliminary remarks

In line with the orientation of this professorship, the typical thesis starts out from a *methodological* issue in econometrics or statistics. Beyond a structured presentation of relevant econometric or statistical methodology it may also include simulations (aka Monte Carlo studies), implementation in statistical software, or a substantial empirical application.

Apart from econometrics and statistics, I also maintain an interest in income distribution and inequality measurement that extends beyond associated statistical methodology.

Also, aspects of stochastic models in economic theory and specifically comparisons of stochastic models via so-called stochastic orders may lead to thesis topics.

Some topics might also be suitable for students with an interest in optimization or, more generally, computation. For example, one could try out various optimizers for certain 'fancy' methods to see if and where they improve upon classical solutions.

See below for further details.

Prerequisites

Please consult first the general rules and requirements of the Faculty of Business and Economics, including information regarding plagiarism.

In general, I can only supervise students who have already acquired at least 30 CPs.

- Students wishing to major in "Quantitative Methods" (old regime) or "Data Science and Computational Economics" (current regime) should have passed the following courses ('equivalent' courses, e.g. at the Faculty of Science, may be acceptable):

    - Econometrics (MSc level) and Fundamentals of Econometric Theory,

    - at least one of Microeconometrics and Statistical Learning or Univariate Time Series Analysis.

    In addition, I strongly recommend to attend further courses in statistics, econometrics, computing, etc. This includes 'Advanced Mathematics for Economics'.

- Students pursuing some other major, or no major, or who are wishing to specialize in other areas, e.g. income distribution, are advised to get in touch at an early stage in order to see whether their background suits the topic. They should have passed

  - Econometrics (MSc level).

  Of course, more is better.

Language

You may write in either German or English. The choice of language does *not* affect the grade. However, good command of the chosen language is expected.

Reproducibility

In economics (and elsewhere) recent concerns about reproducibility of results have led to introduction of new policies at scientific journals regarding provision of data and/or code along with the manuscript. For us, this means that all computational results need to be properly documented. You should therefore expect to submit data – no proprietary formats[1], please use `.txt` or `.csv` – and code in documented form along with your thesis. In the past, I have used CDs for this purpose, but more current storage media are also welcome. For the computations themselves, I encourage the use of R, where reproducibility is relatively straightforward via the `Sweave()` function. There exist more recent tools as well, via the **knitr** or **rmarkdown** packages. The latter do not require knowledge of LaTeX.

Potential topics

- Income distribution and inequality measurement

  From introductory statistics you will probably know the Gini coefficient and the Lorenz curve from a 'descriptive' point of view. There is a huge literature on theory and applications of these and related methods in economics and related fields.

  Tentative topics include aspects of methodology (e.g. estimation from sparse data, such as tables), empirical studies, computational aspects, visualization ('poverty curves', etc.). Potential economic applications include distributional aspects of taxation or the measurement of polarization.

- Size distributions in economics and related areas.

  'Size distributions' refers to distributions of income, wealth, firm sizes, city sizes, actuarial losses, etc. Such data cannot be modeled using familiar distributions such as the normal: they are typically heavily skewed and exhibit 'heavy tails'. Income distributions

---

[1] This includes the MS Excel-specific `.xls`.

were already mentioned above, because there is a large and specialized literature in economics and applied statistics. For other variables such as city sizes there is some overlap with geography, regional science, etc.

Tentative topics include methodological aspects of specific statistical distributions, such as estimation and diagnostics, regression models based on such distributions (e.g., for measuring the influence of demographic variables on income distributions), or empirical applications.

- Regression models for count data.

  A subfield of microeconometrics. Tentative topics include diagnostic checks (e.g. for overdispersion), modified count data models such as zero inflation or hurdle models, models for heavy-tailed data, multivariate extensions.

- Alternative methods of regression and statistical / machine learning.

  Empirical economics, at least in its most applied form, is still dominated by OLS methodology and variations. Here I would be interested in analyses involving more flexible and modern methods, such as quantile regression, robust regression, threshold models, mixture models, and other methods for (potentially) heterogeneous data. This includes methods from the rapidly growing field of statistical learning (called 'machine learning' in computer science), such as boosting, LASSO, regression trees and random forests, etc.

  For example, what if we use a boosting method or random forests in a field of application that traditionally uses logit models for classification?

- Time series analysis.

  Tentative topics include structural change (aka changepoint problems), common features, model diagnostics, forecasting (e.g., new methods, specific applications, 'forecast competitions').

- Stochastic orders in economics.

  This refers to methods for comparing entire distributions. You may have come across concepts such as (first or second-order) stochastic dominance. There are many more! The main applications of these notions are in economic theory (among them decisions under uncertainty or comparisons of income distributions) or public economics (comparisons of tax systems, etc.).

  So this topic isn't about estimation and testing, or data analysis in a wider sense, it may be seen as 'applied probability in business and economics'. It might be suitable for students who like 'pencil and paper' work in microeconomic theory or public economics.

Your own suggestions are also welcome. Please ask at an early stage whether topics are suitable for an M.Sc. thesis.

Data

Typically, I cannot provide data for 'random' topics, so students will need to look into potential data sources for themselves. (My own interests are in methodology, hence I rarely need 'fresh' data.) Sometimes data become available via an internship, which then leads to a thesis topic. In general, obtaining and preprocessing data can be very time-consuming. Students wishing to pursue an empirical thesis are advised to start early.

Some further details

- More and more journals, e.g. *Econometrica*, no longer permit the use of significance stars, and for excellent reasons. Instead, please use standard errors and/or confidence intervals, so that readers can judge for themselves.

- Software and add-on packages should be cited properly. In `R`, try `citation()` or `citation(package = "AER")` for some examples.

- In quantitative methods or economic theory, footnotes are not very common and hence should be avoided. The optimal number is zero, but see p. 2 above.

- The proper handling of references seems to puzzle some students. You will see in the scholarly literature that there is no unified format that is used at all times and in all fields. Solution: pick a format or style that is used in your field of study (e.g., econometrics) and *use it consistently*.

Examples of previous thesis topics

(Titles are given in English, but not all theses were actually written in English.)

Model averaging of binary response models.

Estimating multivariate Poisson mixture regressions, with applications in health economics.

Claim-frequency modeling in non-life insurance pricing: A comparison of generalized linear models with machine-learning methods.

Distributional regression using MultiBoosting and functional gradient descent.

Forecasting economic time series using a hybrid ARIMA-ANN model.

The prediction of SME bankruptcy – An empirical evaluation using machine learning classifiers.

The resource curse: A reanalysis emphasizing outliers and robustness issues.

ARMA subset identification using heuristic optimization.

Easter eggs from Google Trends: Insights from search engine query data for finance and trading.

Early crisis detection using predictive analytics: The case of Africa.

Computation and estimation of dynamic conditional correlation models.

GMM estimation of income distributions from grouped data.

Dating structural breaks in financial time series: A comparison of recent methods.

Changepoint methods in program evaluation.

Structural change in volatility models of Swiss financial time series.

Model choice and robustness in growth regressions.

Forecasting tourism demand for Switzerland.

Redistributive effects of Cantonal income taxes – A simulation study.

Alternative link functions in binary response models.

Observed and unobserved heterogeneity in regressions models for count data.

Heterogeneity in growth regressions.