The Use of Data-driven Transformations and Their Applicability in Small Area Estimation

"Everything should be made as simple as possible, but no simpler"
– Albert Einstein

Representing a relationship between a response variable and a set of covariates is an essential part of the statistical analysis. The linear regression model offers a parsimonious solution to this issue, and hence it is extensively used in nearly all science disciplines. In recent years the linear mixed regression model has become common place in the statistical analysis. Standard statistical techniques for linear and linear mixed regression models are commonly associated with interpretation, estimation, and inference. Numerous assumptions underlying the working model are usually made whenever these models are employed in scientific research. Different options are available to the data analyst when the model assumptions are not met in practice. Researchers could formulate the regression model under alternative and more flexible parametric assumptions. They could also use a regression model that minimizes the use of parametric assumptions or under robust estimation. Another option would be to parsimoniously redesign the model by finding an appropriate transformation such that the model assumptions hold. In general, researchers have been using data transformations as a go-to tool to assist scientific work under the classical and linear mixed regression models instead of developing new theories, applying complex methods or extending software functions. Nevertheless, transformations are often automatically and routinely applied without considering different aspects on their utility. For instance, a standard practice in applied work is to transform the target variable by computing its logarithm. However, this type of transformation does not adjust to the underlying data. Therefore, some research effort has been shifted towards alternative data-driven transformations, which includes a transformation parameter that adjusts to the data. The main contributions of this thesis focus on providing modeling guidelines for practitioners on transformations and on the methodological and practical development of the use of transformations in the context of small area estimation. The proposed approaches are complemented by the development of open source software packages. This aims to close possible gaps between theory and practice.

The literature of transformations in theoretical statistics and practical case studies in different research fields is rich and most relevant results were published during the early 1980s. More sophisticated and complex techniques and tools are available nowadays to the applied statistician as alternatives to using transformations. However, simplification is still a gold nugget in statistical practice, which is often the case when applying suitable transformations within the working model. Some important considerations for using them in linear and linear mixed regression models are still broadly discussed: for example, at which stage of the analysis a transformation should be applied, which transformation is suitable for a specific problem and how the results should be interpreted. For this purpose, the first part of this work proposes some modeling guidelines for practitioners in transformations that seeks to help the researcher to decide if and how a transformation should be applied in practice. An extensive guideline and an overview of different transformations and estimation methods of transformation parameters are presented. It combines a set of pertinent steps, tables, and flowcharts that guide the practitioner through the analysis of transformations in a friendly and practical manner. Additionally, in order to provide an extensive collection of transformations usable in linear regression models and a wide range of estimation methods for the transformation parameter, the package **trafo** is developed and presented as a part of this work. This package

complements and enlarges the methods that exist in **R** so far, and offers a simple, user-friendly framework for selecting a suitable transformation depending on the research purpose.

In the literature, little attention has been paid to the study of techniques of the linear mixed regression model when particularly working with data-driven transformations. This becomes a special challenge for users of small area estimation (SAE) methods, since most commonly used SAE methods are based on the linear mixed regression model which often relies on Gaussian assumptions. In particular, the empirical best predictor is widely used in practice to produce reliable estimates of general indicators for areas with small sample sizes. The issue of data transformations is addressed in the current SAE literature in a fairly ad-hoc manner. Contrary to standard practice in applied work, recent empirical work indicates that using transformations in SAE is not as simple as transforming the target variable by computing its logarithm. The main contributions of this thesis are particularly presented in the second part of the present work, where transformations in the context of SAE are applied and further developed. The study of SAE methods is a research area in official and survey statistics of great practical relevance for national statistical institutes and related organisations. Despite rapid developments in methodology and software, researchers and users would benefit from having practical guidelines for the process of small area estimation. In this thesis a general framework for the production of small area statistics that is governed by the principle of parsimony is proposed. This protocol is based on three stages, namely (i) Specification, (ii) Analysis/Adaptation and (iii) Evaluation. Emphasis is given to the interaction between a user of small area statistics and the statistician in specifying the target geography and parameters in light of the available data. Model-free and model-dependent methods are described with focus on model selection and testing, model diagnostics and adaptations such as use of data transformations. In particular, the use of some adaptations of the working model by using transformations is showed as a part of the (ii) stage. Additionally, the use of data-driven transformations under linear mixed model-based SAE methods is extended; In particular, the estimation method of the transformation parameter under maximum likelihood theory. First, we analyze how the performance of SAE methods are affected by departures from normality and how such transformations can assist with improving the validity of the model assumptions and the precision of small area prediction. In particular, attention has been paid to the estimation of poverty and inequality indicators, due to its important socio-economical relevance and political impact. Second, we adapt the mean squared error estimator to account for the additional uncertainty due to the estimation of transformation parameters. These methodological developments are illustrated using real survey and census data for estimating income deprivation parameters for municipalities in Mexico. Finally, in order to improve some features of existing software packages suitable for the estimation of indicators for small areas, the package **emdi** is developed in this thesis. This package offers a methodological and computational framework for the estimation of regionally disaggregated indicators using SAE methods as well as providing tools for assessing, processing, and presenting the results.

Finally, a discussion of the applicability of transformations is provided in the context of generalized linear models (GLMs). A comparison is made in terms of precision measurements between using count data transformations within the classical regression model and applying GLMs, in particular for the Poisson case. Therefore, some methodological differences are presented and a simulation study is carried out. The learning from this analysis focuses on the relevance of knowing the research purpose and the data scenario in order to choose which methodology should be preferable for any given situation.