Report on the dissertation:

Tomáš Jurczyk: Robustification of statistical and econometrical regression methods

The main contribution of this thesis is the robustification of *Least Weighted Squares* (LWS) in the presence of multicollinearity. Shortly speaking, the author proposes *Ridge Least Weighted Squares* (RLWS) estimator, establishes its basic asymptotic properties (e.g., consistency and break-down point), and investigates its small sample properties by an extensive simulation study.

Considering the existing literature, the proposed methodology is original and I appreciate that the author derives basic asymptotic properties (since this step is often omitted in papers investigating this topic). On the other hand, I disagree with the statement that LTS has not been analyzed in the context of multicollinearity (Chapter 4, line 14–15), see, e.g. [Alfons, A., Croux, C., & Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. The Annals of Applied Statistics, 7(1), 226-248; Kan, B., Alpu, Ö., & Yazıcı, B. (2013). Robust ridge and robust Liu estimator for regression based on the LTS estimator. Journal of Applied Statistics, 40(3), 644-655; Pati, K. D., Adnan, R., & Rasheed, B. A. (2014). Ridge least trimmed squares estimators in presence of multicollinearity and outliers. Nat Sci, 12, 1-8; Amini, M., & Roozbeh, M. (2016). Least trimmed squares ridge estimation in partially linear regression models. Journal of Statistical Computation and Simulation, 1-15.] Interestingly, all of these papers investigate the LTS estimator which can be seen as a special case of the LWS estimator investigated in this dissertation. It should be noted that these papers typically do not investigate consistency or the asymptotic distribution of the proposed estimator.

The dissertation has the usual structure. After an overview of existing literature, the author provides an introduction to linear regression, reviews and illustrates problems caused by contamination and multicollinearity, and introduces some well-known robust estimators (Chapters 1–4). The RLWS estimator is introduced in Chapter 4 and its asymptotic properties (consistency) are established in Chapter 5. Chapter 6 concentrates on the RLTS estimator and investigates its properties by an extensive simulation study. Possible future research directions are discussed in Chapter 7. The appendix contains some previously published theoretical results needed for deriving the theoretical results in Chapter 5.

The most important original results (published in [Jurczyk, T. (2012). Outlier detection under multicollinearity. Journal of Statistical Computation and Simulation, 82(2), 261-278]) are:

1. definition of the RLWS estimator (Chapter 4),
2. proof of the consistency of the RLWS estimator (Chapter 5).

Considering the existing literature on RLTS estimators, it is a pity that Chapter 6 does not concentrate on the more general RLWS estimator.

The dissertation is written in English: the text is reasonably well written and understandable but it is obvious that the author is not a native English speaker (e.g., *Unless* instead of *Although* on page 1, line 6, or *derivation* instead of *derivative* on several occasions).

Questions concerning the proposed methodology:

1. I agree that the RLWS estimator is unique with probability 1 but it is not difficult to construct artificial examples (having probability 0) where this does not hold.

2. Using author's R code on the following example:

```
x=c(1,2,2,2,2)
y=c(0,-1,0,0,1)
w=c(1,1,1,1,0)
LWSridge(cbind(1,x),y,w,delta=0.1)
```

   suggests that the function `LWSridge()` does not provide "consistent" results (sometimes it works but sometimes it returns an error message). What is the reason?

3. It seems that assumptions C2 and NC1 are not satisfied by the RLTS estimator. Are some of the asymptotic results for RLWS valid also for RLTS?

4. How would you calculate standard errors of the RLWS estimates? Is it possible to use the bootstrap approximation although it's based on sampling from a discrete distribution?

Some further comments and questions:

remark 3.2 Do you mean *large change of* Y_i or $\hat{Y}_i$?

page 15 It would be helpful to explain that VIF is connected to "inflation" of some variances.

definition 3.8 Is the minimum calculated over β?

definition 3.9 What exactly is y_0 and z_0?

page 35, line 8 Is the correlation equal to 0 or ρ?

page 36 How do you calculate "identified outliers"? Are these defined as the observations with highest residuals or the "downweighted" observations?

remark 4.3 It is very strange that prior knowledge of β^0 has been used in the simulation study. This may actually invalidate the results of the simulation study because the prior knowledge of the true value of the regression parameter obviously cannot be used in practice.

The results presented in this dissertation are interesting both from practical and theoretical point of view. Compared to existing literature, the author established consistency of the newly defined RLWS. Unfortunately, the assumptions on the weight function (C2) do not allow to apply these results directly to the more popular RLTS estimator that has been already investigated in literature. I appreciate very nice presentation of the effects of multicollinearity and contamination on linear regression (although these effects are well known and thoroughly investigated).

SUMMARY: In my opinion, this dissertation satisfies all necessary requirements given in the rules of the doctoral study programme to justify the award of a PhD degree to Tomáš Jurczyk.

Doc. RNDr. Zdeněk Hlávka, Ph.D.