Outstanding issues in selection of variables and functional forms in multivariable analysis

Willi Sauerbrei[1], Aris Perperoglou[2], Georg Heinze[3] for TG2 of the STRATOS initiative

[1] Institute for Medical Biometry and Statistics, Faculty of Medicine and Medical Center University of Freiburg, Freiburg, Germany

[2] Medical University of Vienna, Austria

[3] Cambridge, UK

Overview

- Multivariable models – preliminaries
- Main aims of TG2
- Issues in variable and function selection
  1. Selection of variables
  2. Selection of functional forms
  3. Combining the two parts
- Outstanding issues
- Assumption
  - Low dimensional data
  - Sample size ‘acceptable’

Building multivariable regression models – some preliminaries

- ‘Reasonable’ model class was chosen
- Comparison of strategies
  - Theory
    - only for limited questions, unrealistic assumptions
  - Examples or simulation
    - Examples based on published data
      - oversimplifies the problem
      - data clean
      - ‚relevant‘ predictors given

      → rigorous pre-selection → what is a full model?
    - Simulations have often weaknesses

… preliminaries continued

- More problems are available,
  - see discussion on initial data analysis in Chatfield (2002) section ‚*Tackling real life statistical problems*‘
  - see also Mallows (1998), The zeroth problem, Am. Stat.
- TG3 – Initial Data Analysis
- Joint talk from Georg Heinze and Marianne Huebner in STRATOS session

➡ Now you should have a statistical analysis plan!

TG2: Selection of variables and functional forms in multivariable analysis

In multivariable analysis, it is common to have a mix of binary, categorical (ordinal or unordered) and continuous variables that may influence an outcome. While TG6 considers the situation where the main task is predicting the outcome as accurately as possible, the main focus of TG2 is to identify influential variables and gain insight into their individual and joint relationship with the outcome. Two of the (interrelated) main challenges are selection of variables for inclusion in a multivariable explanatory model and choice of the functional forms for continuous variables.

[…] The effects of continuous predictors are typically modeled by either categorizing them (which raises such issues as the number of categories, cutpoint values, implausibility of the resulting step-function relationships, local biases, power loss, or invalidity of inference in case of data-dependent cutpoints) or assuming linear relationships with the outcome, possibly after a simple transformation (e.g. logarithmic or quadratic). Often, however, the reasons for choosing such conventional representation of continuous variables are not discussed and the validity of the underlying assumptions is not assessed.

To address these limitations, statisticians have developed flexible modeling techniques based on various types of smoothers, including fractional polynomials and several 'flavors' of splines.

[…] collaborations with other TGs to account for such complexities as missing data, measurement errors, time-varying confounding or issues specific to modeling continuous predictors in survival analyses.

TG2 – Descriptive Models

TG6 - Model for prediction

TG7 – Causal inference

TG2: Part 1 – Selection of variables

- Central issues:
  - Model with focus on prediction or description?
  - To select or not to select (full model)?
  - Which variables to include?
- A large number of methods proposed (for many decades)
- High-dimensional data triggered the development of further proposals
- Many critical issues, state of the art ?

TG2: Overview paper

BioMed Central
The Open Access Publisher

Diagnostic and Prognostic Research

Diagn Progn Res. 2020; 4: 3.
Published online 2020 Apr 2. doi: 10.1186/s41512-020-00074-3

PMCID: PMC7114804
PMID: 32266321

State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues

Willi Sauerbrei,[1] Aris Perperoglou,[2] Matthias Schmid,[3] Michal Abrahamowicz,[4] Heiko Becher,[5] Harald Binder,[1] Daniela Dunkler,[6] Frank E. Harrell, Jr,[7] Patrick Royston,[8] Georg Heinze,[6] and for TG2 of the STRATOS initiative

7 methodological issues identified

Traditional variable selection strategies

- Full model
  - Variance inflation in the case of multicollinearity
- Stepwise procedures
  - forward selection (FS)
  - stepwise selection (StS)
  - backward elimination (BE)
  - Which stop criteria (AIC, BIC, p-value)?

  ➡ has a severe influence on complexity of model selected
- All subset selection
  - which criteria (AIC, BIC)? Or variants of it?

Other procedures

- Procedures based on ‘change-in-estimate’
- Modern variable selection strategies
  - Penalised likelihood
    - Nonnegative garrote
    - (adaptive) lasso
    - Elastic net
    - Smoothly Clipped Absolute Deviation (SCAD)
  - Boosting

… continued

- Resampling-based variable selection procedures
  - Bootstrap inclusion frequencies
    - w/o dependencies among inclusion fractions
  - Stability selection

  ➡ Which type of resampling?

- Bayesian approaches

… continued

- Bias and the role of shrinkage methods

Data dependent model building introduces biases. Several modern selection procedures combine variable selection and shrinkage to correct for it.

Post-estimation shrinkage (2 step approach) can be used for many types of models

- Post-selection inference

Uncertainty caused by model selection is usually ignored. For predictions model averaging approaches have been proposed two decades ago. Many issues unclear and hardly used.

Continuous variables – to categorise or to model?

- Categorisation
  - Avoids some assumptions by introducing others and some severe problems
- Assume linear effect
  - May be wrong
- Modelling nonlinear effects of continuous variables
  - Fractional polynomials (FP)
  - Splines
- ‘spike at zero’ variable - FP procedure proposed

Functional forms:
Models based on cut-points: problems!

- Cut-points are still popular
- Use of cut-points in a model gives a step function
- How many cut-points?
- Where should the cut-points be put?
- Biologically implausible step functions are a poor approximation to the true relationship
- Almost always fits the data less well than a suitable continuous function

- Nevertheless, in many areas still the preferred approach!

Combining variable and function selection

- The multivariable fractional polynomial approach (MFP)
  - Well defined, significance levels are key parameters
- Spline regression
  - Many approaches, hardly any comparisons
  - See Perperoglou talk at IBC 2020

Towards state of the art– research required!

1. Investigation and comparison of the properties of **variable selection strategies**
2. **Comparison of spline procedures** in both univariable and multivariable contexts
3. How to model one or more variables with a ‚**spike-at-zero**‘?
4. Comparison of **multivariable procedures for model and function selection**
5. **Role of shrinkage** to correct for bias introduced by data-dependent modelling
6. Evaluation of new approaches for **post-selection inference**
7. Adaptation of procedures for **very large sample sizes** needed?

Regarding these issues…

- Mathematical theory is unlikely to help
- Simulation studies are key
- However, simulation studies are biased towards the proposed method (Boulesteix et al, 2018)
- Simulation studies are often poorly designed, conducted and reported (Morris et al, 2019)
- Simulation panel of STRATOS works on guidance (see Abrahamowicz talk at IBC 2020)
- Experience from comparative analyses with real data sets
- Translation to level-1 is needed!

Variable selection strategies – some questions

Modern techniques

- Are they useful for low-dimensional situations?
- In which situations do they improve over traditional approaches?
- Is the ‘stability problem’ less severe?
- Do they improve the accuracy of estimates?
- Are there pitfalls in their application for non-expert users?

WHICH modern techniques?

General issues in all studies

- missing data (TG1)
- measurement error (TG4)
- was the study well designed ? (TG5)

Joint work with related STRATOS topic groups

What about state of the art?

State of the art *refers to the* ***highest level of general development****, as of a device, technique, or scientific field achieved at a particular time.*

Wikipedia, 12 June 2017

Conclusion

We are far away from ‘state of the art’ on selection of variables and functional forms

Many more comparisons are urgently needed!

> *‘Exact distributional results are virtually impossible to obtain, even for simplest of common subset selection algorithms’*
>
> *Picard & Cook, JASA, 1984*

➡ **Informative simulation studies are needed!**

… Conclusions

- Member of TG2 identified seven issues
- Other experts may have different experiences and preferences …

  and may raise further issues
- To help deriving evidence-supported guidance, more cooperative and comparative research is needed from experts

Thanks to all members of TG2 !

- Georg Heinze (Austria)
- Aris Perperoglou (U.K.)
- Willi Sauerbrei (Germany)
- Michal Abrahamowicz (Canada)
- Heiko Becher (Germany)
- Harald Binder (Germany)
- Daniela Dunkler (Austria)
- Rolf Groenwold (Netherlands)
- Frank Harrell (U.S.A)
- Nadja Klein (Germany)
- Geraldine Rauch (Germany)
- Patrick Royston (U.K.)
- Matthias Schmid (Germany)

And the early career adjunct members

- Michael Kammer (Vienna, Austria)
- Edwin Kipruto (Freiburg, Germany)
- Christine Wallisch (Vienna, Austria)