

## Variable selection for statistical models: a review and recommendations for the practicing statistician

Georg Heinze, Christine Wallisch and Daniela Dunkler

Section for Clinical Biometrics  
Center for Medical Statistics, Informatics and  
Intelligent Systems  
Medical University of Vienna  
Vienna, Austria

for Topic Group 2 of the STRATOS initiative  
([www.stratos-initiative.org](http://www.stratos-initiative.org))  
*E-mail:* [georg.heinze@meduniwien.ac.at](mailto:georg.heinze@meduniwien.ac.at)

Statistical models are important tools in empirical medical research. They facilitate individualized outcome prognostication conditional on covariates as well as adjustments of estimated effects of covariates on the outcome. Theory of statistical models is well-established if the set of covariates to consider is fixed and small, such that we can assume that effect estimates are unbiased and the usual methods for confidence interval estimation are valid. In routine work, however, it is not known a priori which covariates should be included in a model, and often we are confronted with the number of candidate variables in the range 10-25. This number is often too large to be considered in a statistical model.

In recent decades many statisticians have extensively studied variable selection procedures for various purposes, e.g., for adjusting the effect of a risk factor of interest for confounders or other covariates, for hypothesis testing, or for deriving multivariable prediction models. It has turned out that no selection procedure is generally superior to other procedures and there is no generally accepted state of the art for variable selection [1]. Unfortunately, in medical papers it is still not uncommon to use univariable selection as a screening approach to eliminate non-significant variables and use the remaining variables to build the multivariable model. This approach has severe weaknesses. We will provide an overview of variable selection methods which are based on

- a) significance or information criteria, [2; Ch. 2]
- b) penalized likelihood, [3]
- c) the change-in-estimate criterion, [4]
- d) background knowledge, [5] or
- e) combinations thereof. [6]

These methods were usually developed in the context of a linear regression model and then transferred to more general models like generalized linear models or models for censored survival data.

In this half-day workshop, we will exemplify applications of variable selection using scientific questions and data from real medical studies with different research questions focusing on descriptive models and transparent prediction models. Data of these studies are publicly available, and their analysis will be discussed by means of worked exercises with accompanying R notebooks. We will also interactively present a simulation study to investigate implications of variable selection, e.g., on uncertainty and stability of the final model [7,8], on bias and variability of regression coefficients [9], and on the validity of confidence intervals [10].

We will give pragmatic recommendations for the practitioner by suggesting typical steps to be done when variable selection is considered. We give guidance on how to pre-select candidate covariates, how to choose an appropriate variable selection method, and how to report the final model and its

stability in scientific reports [11,12]. These recommendations are based on data settings with a mix of 5-25 continuous and categorical covariates that are moderately correlated ( $r < 0.8$ ). We also discuss some open issues that still need further investigation [1].

We will mix visual presentations with check-up questions to the audience and will demonstrate worked exercises interactively in R-Studio. Participants can follow these analyses with their own notebook, but it is not required to bring a notebook to attend and follow this course.

#### References:

- [1] Sauerbrei W, Perperoglou A, Schmid M, Abrahamowicz M, Becher H, Binder H, Dunkler D, Harrell Jr FE, Royston P, Heinze G, for TG2 of the STRATOS initiative. State of the art in selection of variables and functional forms in multivariable analysis – outstanding issues. *Diagnostic and Prognostic Research* 4:3, 2020.
- [2] Royston P, Sauerbrei W. *Multivariable Model-Building. A pragmatic approach to regression analysis based on fractional polynomials for modeling continuous variables*. Wiley, Chichester, 2008
- [3] Tibshirani R. Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B* 58: 267–288, 1996
- [4] Mickey RM, Greenland S. The impact of confounder selection criteria on effect estimation. *American Journal of Epidemiology* 129: 125–137, 1993
- [5] VanderWeele TJ, Shpitser I. A new criterion for confounder selection. *Biometrics* 67: 1406–1413, 2011
- [6] Dunkler D, Plischke M, Leffondré K, Heinze G. Augmented backward elimination: A pragmatic and purposeful way to develop statistical models. *PLoS One* 9(11): e113677, 2014
- [7] Buckland ST, Burnham KP, Augustin NH. Model selection: An integral part of inference. *Biometrics* 53: 603-618, 1997
- [8] Sauerbrei W, Schumacher M. A bootstrap resampling procedure for model building: application to the Cox regression model. *Statistics in Medicine* 11: 2093–2109, 1992
- [9] Steyerberg EW, Schemper M, Harrell FE. Logistic regression modeling and the number of events per variable: selection bias dominates. *Journal of Clinical Epidemiology* 64(12), 1464-5, 2011
- [10] Austin PC. Using the bootstrap to improve estimation and confidence intervals for regression coefficients selected using backwards variable elimination. *Statistics in Medicine* 27, 3286-3300, 2008
- [11] Heinze G, Wallisch C, Dunkler D. Variable selection – a review and recommendations for the practicing statistician. *Biometrical Journal* 60, 431-449, 2018
- [12] Wallisch C, Dunkler D, Rauch G, de Bin R, Heinze G. Selection of variables for multivariable models: Opportunities and limitations in quantifying model stability by resampling. *Statistics in Medicine* 40:369-381, 2021

## Course instructors

Daniela Dunkler	<a href="https://www.meduniwien.ac.at/researcher/daniela_dunkler">https://www.meduniwien.ac.at/researcher/daniela_dunkler</a>
Georg Heinze	<a href="https://www.meduniwien.ac.at/researcher/georg_heinze">https://www.meduniwien.ac.at/researcher/georg_heinze</a>
Christine Wallisch	<a href="https://www.meduniwien.ac.at/researcher/christine_wallisch">https://www.meduniwien.ac.at/researcher/christine_wallisch</a>

## Agenda

### First session (2pm – approx. 3.30pm)

I-1 Philosophy of parsimonious modeling (Christine Wallisch)

VS Handout 1.pdf

I-2 Toolbox (Georg Heinze)

VS Handout 2.pdf

I-3 Algorithms of variable selection (Georg Heinze)

VS Handout 3.pdf

Break

### Second session (approx. 4pm – 5.30pm)

II-4 Consequences of variable selection – a simulation study (Georg Heinze)

VS Handout 4.pdf

+ interactive live-demonstration 'Visualization of simulation results: Comparison of variable selection methods' (*shiny-app not included in material*)

II-5 Case studies (Christine Wallisch)

VS Handout 5.1.pdf (Slides)

VS Handout 5.2.pdf (compiled R markdown)

VS Markdown ad 5.2.Rmd (R markdown file for reproducibility)

+ interactive RStudio session using HTML Vignettes (*not included in material*)

II-6 Towards recommendations (Georg Heinze)

VS Handout 6.pdf

## References (from slides)

- Breiman, L. (2001a). Statistical modeling: the two cultures. *Statistical Science* **16**, 199-231.
- Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model selection: an integral part of inference. *Biometrics* **53**, 603-618.
- Burnham, K. P., and Anderson, D. R. (2002). Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. *Springer*, New York.
- Cowling, T.E., Cromwell, D.A., Sharples, L.D., Van Der Meulen, J., (2020). A novel approach selected small sets of diagnosis codes with high prediction performance in large healthcare datasets. *Journal of Clinical Epidemiology* **128**, 20–28, doi.org/10.1016/j.jclinepi.2020.08.001
- Cox, D. R., and Hinkley, D. V. (1979). *Theoretical Statistics*, 1st edition. Boca Raton: *Chapman and Hall/CRC*.
- Desboulets, L., (2018). A Review on Variable Selection in Regression Analysis. *Econometrics* **6**, 45.. doi:10.3390/econometrics6040045
- Dunkler, D., Sauerbrei, W., and Heinze, G. (2016). Global, parameterwise and joint shrinkage factor estimation. *Journal of Statistical Software* **69**, 1-19.
- Dunkler, D., Plischke, M., Leffondré, K., and Heinze, G. (2014). Augmented backward elimination: A pragmatic and purposeful way to develop statistical models. *PLoS ONE* **9**, doi: 10.1371/journal.pone.0113677.
- Eichinger, S., Heinze, G., Jandek, L.M., Kyrle, P.A., (2010). Risk Assessment of Recurrence in Patients With Unprovoked Deep Vein Thrombosis or Pulmonary Embolism. *Circulation* **121**, 1630–1636., doi:10.1161/circulationaha.109.925214
- Glymour, M.M., Weuve, J. & Chen, J.T. (2008). Methodological Challenges in Causal Research on Racial and Ethnic Patterns of Cognitive Trajectories: Measurement, Selection, and Bias. *Neuropsychol Rev* **18**, 194–213. doi.org/10.1007/s11065-008-9066-x.
- Good, D.M., Zürgbilg, P., Argilés, À., Bauer, H.W., Behrens, G., Coon, J.J., Dakna, M., Decramer, S., Delles, C., Dominiczak, A.F., Ehrlich, J.H.H., Eitner, F., Fliser, D., Frommberger, M., Ganser, A., Girolami, M.A., Golovko, I., Gwinner, W., Haubitz, M., Herget-Rosenthal, S., Jankowski, J., Jahn, H., Jerums, G., Julian, B.A., Kellmann, M., Kliem, V., Kolch, W., Krolewski, A.S., Luppi, M., Massy, Z., Melter, M., Neusüss, C., Novak, J., Peter, K., Rossing, K., Rupprecht, H., Schanstra, J.P., Schiffer, E., Stolzenburg, J.-U., Tarnow, L., Theodorescu, D., Thongboonkerd, V., Vanholder, R., Weissinger, E.M., Mischak, H., Schmitt-Kopplin, P., (2010). Naturally occurring human urinary peptides for use in diagnosis of chronic kidney disease. *Molecular & cellular proteomics : MCP*, **9**(11), 2424–2437. doi.org/10.1074/mcp.M110.001917
- Greenland, S. (2000). Principles of multilevel modelling. *International Journal of Epidemiology* **29**, 158-167.
- Greenland S. (1997). Second-stage least squares versus penalized quasi-likelihood for fitting hierarchical models in epidemiologic analyses. *Stat Med*, Mar **15**;16(5):515-26. doi: 10.1002/(sici)1097-0258(19970315)16:5<515::aid-sim425>3.0.co; 2-v. PMID: 9089960.
- Greenland, S., Daniel, R., Pearce, N. (2016). Outcome modelling strategies in epidemiology: traditional methods and basic alternatives, *International Journal of Epidemiology*, **45**(2):565–575. doi.org/10.1093/ije/dyw040
- Hafermann, L., Becher, H., Herrmann, C., Klein, N., Heinze, G., Rauch, G. (2021). Statistical Model Building: Background "Knowledge" Based on Inappropriate Preselection causes Misspecification.
- Harrell, F. E. (2001). *Regression Modeling Strategies. With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer, 1<sup>st</sup> Edition, New York.
- Harrell, F. E. (2015). *Regression Modeling Strategies. With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer, New York, Berlin, Heidelberg.
- Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edition. Springer, New York.
- Heinze, G., and Dunkler, D. (2017). Five myths about variable selection. *Transplant International* **30**, 6-10.
- Heinze, G., Wallisch, C., Dunkler, D. (2018). Variable selection - A review and recommendations for the practicing statistician. *Biometrical Journal* **60**, 431–449. doi:10.1002/bimj.201700067
- Hosmer, D. W., Lemeshow, S., and May, S. (2011). *Applied Survival Analysis: Regression Modeling of Time to Event Data*, 2nd edition. Wiley, Hoboken, NJ.
- Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied Logistic Regression*, 3rd edition. Wiley, Hoboken, NJ.
- Johnson, R. W. (1996). Fitting percentage of body fat to simple body measurements. *Journal of Statistics Education* **4**:1.
- Kammer, M., Dunkler, D., Michiels, S., Heinze, G. (2021). Evaluating methods for Lasso selective inference in biomedical research by a comparative simulation study. *Preprint on arXiv.org*, arxiv.org/abs/2005.07484.
- Lee, P. H. (2014). Is a cutoff of 10% appropriate for the change-in-estimate criterion of confounder identification? *Journal of Epidemiology* **24**, 161-167.
- Lu, Z., Lou, W. (2021) "Bayesian approaches to variable selection: a comparative study from practical perspectives: " The International Journal of Biostatistics, vol. , no. , 2021, pp. 20200130. https://doi.org/10.1515/ijb-2020-0130
- Maldonado, G., and Greenland, S. (1993). Simulation study of confounder-selection strategies. *American Journal of Epidemiology* **138**, 923-936.
- Meinshausen, N., and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society Series B-Statistical Methodology* **72**, 417-473.
- Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika* **82**, 669–688.
- Polterauer, S., Grimm, C., Hofstetter, G., Concini, N., Natter, C., Sturza, A., Pötter, R., Marth, C., Reinthaller, A., Heinze, G. (2012). Nomogram prediction for overall survival of patients diagnosed with cervical cancer. *British Journal of Cancer* **107**, 918–924. doi:10.1038/bjc.2012.340

- Riley, R., Snell, K.I., Martin, G., Whittle, R., Archer, L., Sperrin, M., Collins, G. (2020). Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. *Journal of Clinical Epidemiology*, **132**, 88 - 96.
- Royston, P., and Sauerbrei, W. (2008). *Multivariable Model-building. A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials For Modelling Continuous Variables*. John Wiley & Sons, Ltd, Chichester, UK.
- Rubin, D. B. (2009). Author's reply (to Judea Pearl's and Arvid Sjölander's letters to the editor). *Statistics in Medicine* **28**, 1420–1423.
- Sauerbrei, W. (1999). The use of resampling methods to simplify regression models in medical statistics. *Journal of the Royal Statistical Society Series C-Applied Statistics* **48**, 313-329.
- Sauerbrei, W., and Schumacher, M. (1992). A bootstrap resampling procedure for model building: Application to the Cox regression model. *Statistics in Medicine* **11**, 2093-2109.
- Sauerbrei, W., Perperoglou, A., Schmid, M., Abrahamowicz, M., Becher, H., Binder, H., Dunkler, D., Harrell, F.E., Royston, P., Heinze, G., (2020). State of the art in selection of variables and functional forms in multivariable analysis—outstanding issues. *Diagnostic and Prognostic Research* **4**. doi:10.1186/s41512-020-00074-3
- Sheppard, J. P., Stevens, R., Gill, P., Martin, U., Godwin, M., Hanley, J., Heneghan, C., Hobbs, F. D., Mant, J., McKinstry, B., Myers, M., Nunan, D., Ward, A., Williams, B., & McManus, R. J. (2016). Predicting Out-of-Office Blood Pressure in the Clinic (PROOF-BP): Derivation and Validation of a Tool to Improve the Accuracy of Blood Pressure Measurement in Clinical Practice. *Hypertension (Dallas, Tex. : 1979)*, **67**(5), 941–950. doi.org/10.1161/HYPERTENSIONAHA.115.07108
- Shmueli, G. (2010). To explain or to predict? *Statistical Science* **25**, 289-310.
- Sinkovec, H., Heinze, G., Blagus, R., Geroldinger, A. (2021). To tune or not to tune, a case study of ridge logistic regression in small or sparse datasets. Preprint on *arxiv.org*. arxiv.org/abs/2101.11230
- Steyerberg, E. (2009). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Springer, New York.
- Sullivan, S.G., Greenland, S. (2013). Bayesian regression in SAS software. *International Journal of Epidemiology* **42**, 308–317. doi:10.1093/ije/dys213
- Sun, G.-W., Shook, T. L., and Kay, G. L. (1996). Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *Journal of Clinical Epidemiology* **49**, 907-916.
- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B-Methodological* **58**, 267-288.
- Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**, 273–282. doi:10.1111/j.1467-9868.2011.00771.x
- Van Calster, B., van Smeden, M., De Cock, B., & Steyerberg, E. W. (2020). Regression shrinkage methods for clinical prediction models do not guarantee improved performance: Simulation study. *Statistical methods in medical research*, **29**(11), 3166–3178. https://doi.org/10.1177/0962280220921415
- Van Der Ploeg, T., Austin, P.C., Steyerberg, E.W. (2014). Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology* **14**, 137. doi:10.1186/1471-2288-14-137
- Van Smeden, M., De Groot, J.A.H., Moons, K.G.M., Collins, G.S., Altman, D.G., Eijkemans, M.J.C., Reitsma, J.B. (2016). No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology* **16**. doi:10.1186/s12874-016-0267-3
- VanderWeele, T. J., and Shpitser, I. (2011). A new criterion for confounder selection. *Biometrics* **67**, 1406-1413. doi:10.1111/j.1541-0420.2011.01619.x
- Verweij, P. J. M., Van Houwelingen, H. C., (1993). Cross-validation in survival analysis. *Statistics in Medicine*, **12**: 2305-2314. doi.org/10.1002/sim.4780122407
- Wallisch, C., Dunkler, D., Rauch, G., Bin, R., Heinze, G., (2021). Selection of variables for multivariable models: Opportunities and limitations in quantifying model stability by resampling. *Statistics in Medicine* **40**, 369–381. doi:10.1002/sim.8779
- Witte, J., Didelez, V. (2019). Covariate selection strategies for causal inference: Classification and comparison. *Biometrical Journal* **61**, 1270–1289. doi:10.1002/bimj.201700294
- Wynants, L., Van Calster, B., Collins, G.S., Riley, R.D., Heinze, G., Schuit, E., Bonten, M.M.J., Dahly, D.L., Damen, J.A., Debray, T.P.A., De Jong, V.M.T., De Vos, M., Dhiman, P., Haller, M.C., Harhay, M.O., Henckaerts, L., Heus, P., Kammer, M., Kreuzberger, N., Lohmann, A., Luijken, K., Ma, J., Martin, G.P., McIernon, D.J., Andaur Navarro, C.L., Reitsma, J.B., Sergeant, J.C., Shi, C., Skoetz, N., Smits, L.J.M., Snell, K.I.E., Sperrin, M., Spijker, R., Steyerberg, E.W., Takada, T., Tzoulaki, I., Van Kuijk, S.M.J., Van Bussel, B.C.T., Van Der Horst, I.C.C., Van Royen, F.S., Verbakel, J.Y., Wallisch, C., Wilkinson, J., Wolff, R., Hooft, L., Moons, K.G.M., Van Smeden, M. (2020). Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal *BMJ* 2020; **369** :m1328 doi:10.1136/bmj.m1328

ROeS conference 2021 pre-conference workshop

# Variable selection – a review and recommendations for the practicing statistician

Georg Heinze, Christine Wallisch & Daniela Dunkler  
Medical University of Vienna  
CeMSIIS – Section for Clinical Biometrics  
For TG2 of the STRATOS initiative



[georg.heinze@meduniwien.ac.at](mailto:georg.heinze@meduniwien.ac.at), [christine.wallisch@meduniwien.ac.at](mailto:christine.wallisch@meduniwien.ac.at), [daniela.dunkler@meduniwien.ac.at](mailto:daniela.dunkler@meduniwien.ac.at)

## Aims of the lecture

- To explain aspects of **variable selection in multivariable regression** analyses of **observational studies**.
- To review **different variable selection strategies and modeling philosophies**.
- To encourage investigations of **model instability** induced by variable selection.
- To illustrate the urgent **need for background knowledge** in statistical modeling.

## Agenda

- Part I-1: Philosophy
- Part I-2: Prerequisites
- Part I-3: Variable selection methods and strategies

*Break*

- Part II-1: Consequences of variable selection
- Part II-2: Case studies
- Part II-3: Recommendations



## PART I-1: PHILOSOPHY

Magritte, Ockham, Einstein

## What is this?



## What is this?



„This is not a pipe“  
René Magritte, 1928-29



## Aim of this part

- To understand how models simplify and approximate reality.

## What do we mean by a statistical model?

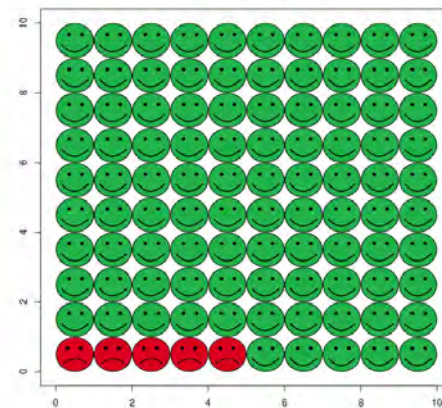
- *A set of probability distributions on the sample space  $\mathcal{S}$ .*  
(e.g. Cox and Hinkley, 1974)
- *Statistical models summarize patterns of the data available for analysis.*  
(Steyerberg, 2009)
- *A powerful tool for developing and testing theories by way of causal explanation, prediction, and description.*  
(Shmueli, 2010)
- *A simplification or approximation of reality.*  
(Burnham, Anderson, 2002)
- *A model represents, often in considerably idealized form, the data-generating process.* (Wikipedia)

## What do we mean by a statistical model?

- *Statistical models are simple mathematical rules derived from empirical data describing the association between an outcome and several explanatory variables.* (Dunkler et al, 2014)
- They should be **valid**: provide predictions with acceptable accuracy.
- They should be **practically useful**: allow conclusions such as ‘how large is the expected difference in outcome if one of the explanatory variables differs by one unit’.
- They should be **robust**.

## What are typical components of a statistical model?

Based on the provided information, the risk to suffer from cardiovascular disease within the next 5 years is 4.8%.  
One can expect that 5 out of 100 individuals with these characteristics will develop cardiovascular disease within the next 5 years.



Prediction of 5-year  
general cardiovascular risk

for individuals without  
any prior cardiovascular disease  
between 30 to 74 years.

<https://cvdrisk.shinyapps.io/english>

## What can we learn from this model?

- **Prediction**

*Risk Score = 4.8%.*

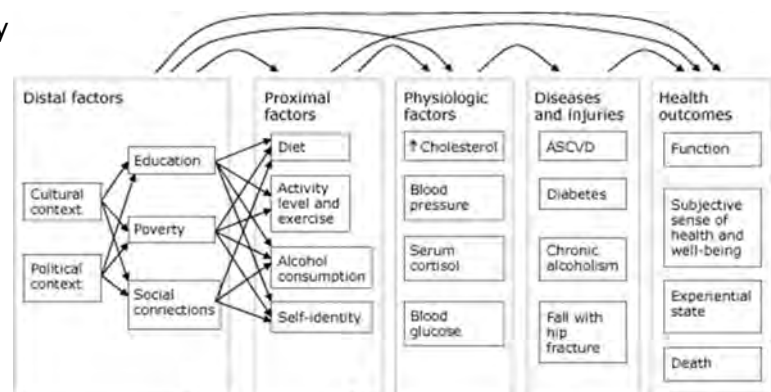
*Means 5 of 100 people with this level of risk will have a cardiovascular event in the next 5 years.*

- **Explanation**

*A person with a controlled systolic blood pressure with 120 mmHg has 1.3-times the risk of a person with natural systolic blood pressure of 120 mmHg.*

## Why multivariable modeling?

- Disease causation is usually multifactorial.
- Co-prognostic variables can only be identified in a multivariable context.



(from [http://www.cdc.gov/pcd/issues/2010/jul/10\\_0005.htm](http://www.cdc.gov/pcd/issues/2010/jul/10_0005.htm))

## Purposes of multivariable models

- Prediction of an outcome of interest
- Identification of 'important' predictors
- Understanding the effects of predictors ('explanatory')
- Adjustment for predictors uncontrollable by experimental design
- Stratification by risk

(Royston & Sauerbrei, 2008)

## To Explain or to Predict?

- Shmueli (2010):
  - **Prediction models**
  - **Descriptive models**
  - **Explanatory models**

## To Explain or to Predict?

- **Prediction models**
  - Interest in accurate predictions for future application.
  - No concern about causality and confounding (association).
  - Diagnostic and prognostic prediction models.
- **Aims of prediction:**
  - Transparent: formula-based predictions can be explained as/decomposed in contributions of X's
  - Simple: model is more easily applicable with few variables
  - Misspecification may lead to locally biased predictions

## To Explain or to Predict?

- **Descriptive models**
  - Capture the data structure parsimoniously:  
which variables are associated with the outcome and how?
  - Often useful transparent prediction models,  
in special cases even causal conclusions possible
- **Aims of description:**
  - Just X and Y: understand how Y is associated with X's
  - Simple: make general, widely valid statements about these associations
  - Misspecification ,by intention'

## To Explain or to Predict?

- **Explanatory models**
  - Interest in causal contrasts (e.g., coefficients)
  - Often achieved by counterfactual prediction
  - Confounder selection
- **Aims of explanation** (causal inference):
  - Interest in effect of an intervention on an outcome
  - Main concern: correct adjustment for confounders
  - Misspecification leads to biased effect estimate
  - Simplicity not ultimately needed; may reduce variance

## Model building depends on study aim

1. The model is predefined. Estimate parameters and check assumptions. (Randomized trial.)
2. Develop a good predictor. Number of variables should be small.
3. Develop a good predictor. No limits in model complexity.
4. Assess the effect of a new factor of interest, adjusting for established factors.
5. Assess the effect of a new factor of interest, adjusting for confounding factors selected by data analysis.
6. Hypothesis generation of possible associations of factors with outcome in studies with many covariates.

Data-driven!



(Royston & Sauerbrei, 2008)

## Is there a true model?

A 'true model' = a 'true data generating mechanism'.

### Pro:

- Aristotle: *'Nature operates in the shortest way possible.'*
- Newton: *'We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances.'*

## Is there a true model?

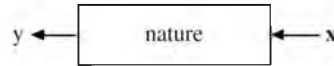
A 'true model' = a 'true data generating mechanism'.

### Contra:

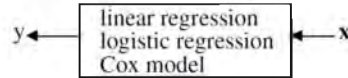
- *'We do not accept the notion that there is a simple "true model" in the biological sciences.'* (Burnham & Anderson, 2002)
- *'We recognize that true models do not exist... A model will only reflect underlying patterns, and hence should not be confused with reality.'* (Steyerberg, 2009)
- *'I started reading Annals of Statistics, and was bemused: Every article started with „Assume that the data are generated by the following model: ..." followed by mathematics exploring inference, hypothesis testing and asymptotics.'* (Breiman, 2001)
- *'All models are wrong, but some are useful.'* (Box)

## Do we need statistical models at all?

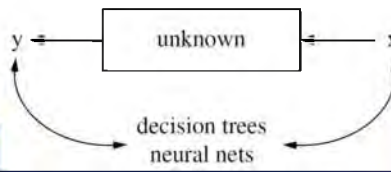
- Statistics starts with data. These data are 'generated' inside a black box by nature.



- Statistical culture I:* Assume a stochastic data model for the inside of the box.

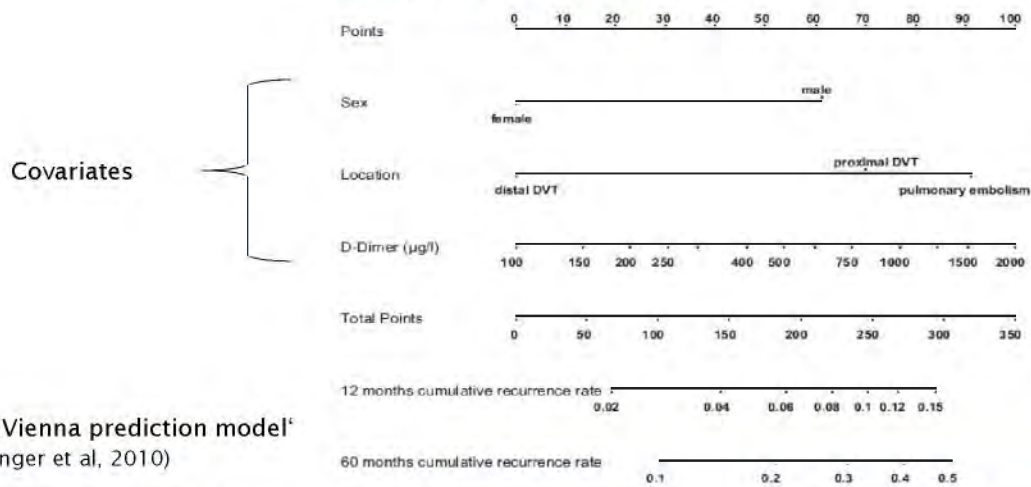


- Statistical culture II:* The inside of the box is complex and unknown. Find a function  $f(X)$  - an algorithm - that operates on  $X$  to predict the responses  $Y$ .



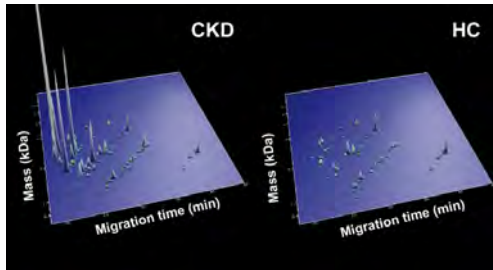
(Breimann, 2001)

## Example I: Prediction of recurrence of venous thromboembolism

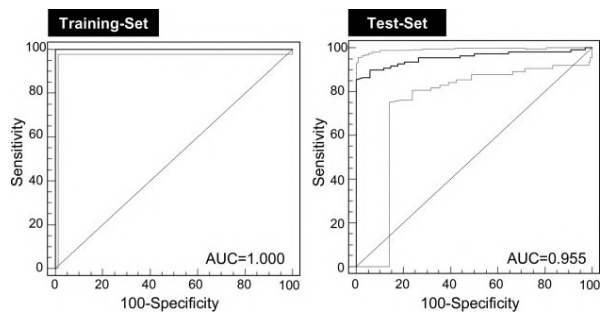




## Example II: Urine-proteomic predictor of incidence of early chronic kidney disease



Support Vector Machine



CKD273 predictor  
(Good et al, 2010)

## William of Ockham

- 14<sup>th</sup> century logician and Franciscan friar:  
*'Pluralitas non est ponenda sine neccesitate.'*  
(Entities should not be multiplied unnecessarily.)
- When you have 2 competing theories that make exactly the same predictions, the simpler one is the better.
- If you have 2 equally likely solutions to a problem, choose the simplest.
- The explanation requiring the fewest assumptions is most likely to be correct.
- *'Simplicity is the ultimate sophistication.'* (Leonardo da Vinci)
- *'Everything should be made as simple as possible, but not simpler.'* (~Einstein)

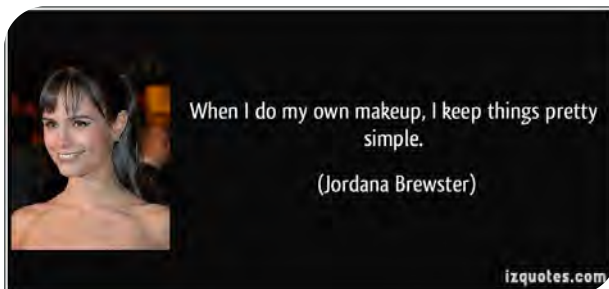


## Summary

- Models are not reality.
- There is no such thing as a 'true model'.  
=> There is not a single model that will ultimately explain data generation.
- Models can be useful: for pure prediction or  
for understanding multidimensional association.
- If two models have the same explanatory power, we should prefer the simpler one.
- Complex models can be more accurate than simple ones, but are often less useful  
(for description or prediction).

## Focus of this presentation

- Methods and consequences of variable selection





## Part I-2: Prerequisites

MEDICAL UNIVERSITY OF VIENNA

STRATOS INITIATIVE

Georg Heinze, Christine Wallisch, Daniela Dunkler  
CeMSIIS - Section for Clinical Biometrics  
2

Part I-

## Aim

- To explain statistical/mathematical prerequisites that are useful in variable selection.
- Focusing on a descriptive or transparent prediction research aim, seasoned with a knife tip of causal inference.

MEDICAL UNIVERSITY OF VIENNA

STRATOS INITIATIVE

Georg Heinze, Christine Wallisch, Daniela Dunkler  
CeMSIIS - Section for Clinical Biometrics

Part I-2 2

## Statistical prerequisites

The diagram 'Statistical prerequisites' features a central point with several branches, each in a different colored box:

- Types of models by distribution of error** (blue box)
- Assumptions of models** (red box)
- Hypothesis tests: Likelihood ratio, Score, Wald** (orange box)
- Model estimation: maximum likelihood** (purple box)
- Likelihood and information-theoretic measures** (orange box)
- Shrinkage** (blue box)
- Resampling techniques** (purple box)
- AIC and AICc** (orange box)
- Penalized likelihood** (purple box)
- Confounding** (green box)
- Prior knowledge** (black box)
- Bias-variance tradeoff** (blue box)
- Change-in-estimate criterion** (orange box)

MEDICAL UNIVERSITY OF VIENNA

Georg Heinze, Christine Wallisch, Daniela Dunkler  
 CeMSIS - Section for Clinical Biometrics

Part I-2 3

## Preselection of variables

- Background knowledge/domain expertise!
- Chronology
- Costs of collecting measurements
- Availability at time of model use
- Quality (measurement errors)
- Confounder criteria

}

Discussion with domain expert!

- Availability in data set (missing values)
- Variability (rare categories)
- Preselection = assuming no other variables important!

MEDICAL UNIVERSITY OF VIENNA

Georg Heinze, Christine Wallisch, Daniela Dunkler  
 CeMSIS - Section for Clinical Biometrics

Part I-2 4

## What models do we typically see?

### Linear model

- $Y = \beta_0 + \beta_1 X_1 + \dots + \beta_K X_k + \epsilon = X\beta + \epsilon$
- $\epsilon \sim N(0, \sigma)$

### Logistic model

- $\Pr(Y = 1) = \text{expit}(\beta_0 + \beta_1 X_1 + \dots + \beta_K X_k)$   
 $= \exp(X\beta) / [1 + \exp(X\beta)]$

### Cox model

- $h(X, t) = h_0(t) \exp(\beta_1 X_1 + \dots + \beta_K X_k) = h_0(t) \exp(X\beta)$

## Common assumptions

### Linearity: linear combination of variables

- (Relaxation: splines, fractional polynomials, GAMs)

### Additivity: sum of effects

- (Relaxation: include interactions, power functions, etc.)

## Interpretation of regression coefficients



- Adjusted effect of  $X_k$ :
- Expected difference in outcome, if  $X_k$  differs by 1 unit and all other  $X$ 's constant.
- $\beta_k$  measures the 'independent' effect of  $X_k$ .
- Fundamentally different in different models!

## Interpretation of regression coefficients

- Consider the following models to explain %body fat:

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	76.65092	9.97648	7.68	<.0001
height_cm	Height in cm	1	-0.58611	0.06204	-9.45	<.0001
weight_kg	Weight in kg	1	0.58177	0.03368	17.28	<.0001

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	-30.36370	11.43150	-2.66	0.0084
abdomen	Abdomen circumference	1	0.91008	0.07137	12.75	<.0001
weight_kg	Weight in kg	1	-0.21541	0.06778	-3.18	0.0017
height_cm	Height in cm	1	-0.09593	0.06171	-1.55	0.1213

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	-14.89166	2.76160	-5.39	<.0001
weight_kg	Weight in kg	1	0.41950	0.03371	12.44	<.0001

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	-47.65873	2.63417	-18.09	<.0001
abdomen	Abdomen circumference	1	0.97919	0.05599	17.49	<.0001
weight_kg	Weight in kg	1	-0.29219	0.04655	-6.28	<.0001

## Provided information versus desired knowledge

- Information provided by the data:
  - Number of independent observations  $N$
  - Number of events  $E$   
(logistic:  $\min(\#events, \#non-events)$ , Cox:  $\#events$ )
- Amount of knowledge desired:
  - Number of unknown regression coefficients ( $K$ )
- Summarized by 'events per variable'  $EPV = E/K$ ,  $NPV = N/K$ .
- Often cited minimum  $EPV = 10$  is questionable.

## Events Per Variable (EPV)

- $EPV \geq 10$  (Harrell 2001, p. 61) ... or  $EPV \geq 15$  (Harrell 2015)
  - Number of candidate variables, not variables in the final model.
  - Should be considered as lower bound!
  - See also van Smeden et al (2016):  
,No rationale for 1 variable per 10 events...'
- Non-linearity, interactions, etc.  $\rightarrow EPV \uparrow$ .
- Prediction  $\rightarrow EPV \uparrow$  (logistic regression  $EPV$  20–50).
- Modern modeling techniques (random forests, neural networks, support vector machines)  $\rightarrow$  10 times  $EPV$  compared to logistic regression  $\rightarrow EPV \uparrow \uparrow$   
(van der Ploeg et al. 2014).

## Testing coefficients of models

- Consider two hierarchically nested models ( $M_2$  nested in  $M_1$ ; in  $M_2$  some  $\beta = 0$ )
- Wald tests: use only  $M_1 \rightarrow$  step down
- Scores tests: use only  $M_2 \rightarrow$  step up
- Likelihood ratio test: compare  $M_1$  and  $M_2$ ; considered the most precise test.



Ronald A. Fisher  
in 1913



Abraham Wald,  
1902-1950

## Testing between models

- What does it mean to test models?
  - OK if the test is 'prespecified' – rarely done in practice.
  - Not informative if models result from earlier testing (iterated testing: tests on 'generated' hypotheses).
- Consequence:
  - 'Tests' are interpretable if a few, pre-specified working models are compared.
  - We cannot trust the p-values from selected models!
- Modeling and hypothesis testing – two hostile brothers?





## Akaike information criterion

- Akaike showed that for model selection we need to maximize the ‘cross-validated’ expectation of  $\log L$  across several competitive models:

$$E_{test}E_{train}[\log L(x_{test}|\hat{\beta}_{train})]$$

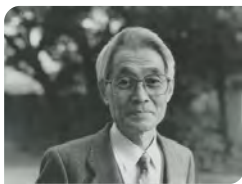
Model developed on  $x_{train}$ ,  
Evaluated on  $x_{test}$ .

- This can be approximated by

$$\log L(x_{train}|\hat{\beta}_{train}) - K$$

Model developed on  $x_{train}$ ,  
Evaluated on  $x_{train}$ .

- He defined  $AIC = -2 \log L(x_{train}|\hat{\beta}_{train}) + 2K$ .



Hirotumi Akaike, 1927-2009,  
(from <http://andrewgelman.com>)

$K$  ... number of parameters

## The value of AIC

- We can compare two non-hierarchical models.
- We can compare several models.
- Hierarchical models: corresponding p-values

Degrees of freedom difference	Equivalent p-value in LR test
1	0.157
2	0.135
3	0.117
4	0.092

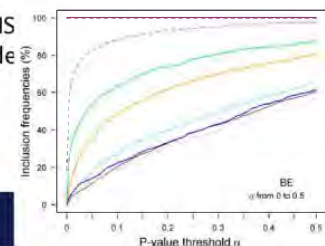
- General:  $1 - \text{pchisq}(2 * df, df)$

## Schwarz's Bayesian Information Criterion (BIC)

- Defined as  $BIC = -2 \log L + \log(N)K$
- If the 'true' model is among the candidate models, then BIC will select the true model as  $N \rightarrow \infty$  (consistent model selection)
- For Cox or logistic models,  $N'$  is the number of events, or  $\min(\text{events}, \text{non-events})$
- More stringent selection for large N than for small N
- Compute equivalent significance level in R by `1-pchisq(log(N)*K, K)`
- For  $K=1, N=100$ : equivalent to  $\alpha = 0.032$
- $\rightarrow$  AIC selects more variables than BIC

## Resampling to quantify model stability

- **Model selection frequency (MSF):** how likely is selection of a model?  
- the 'final' model  
- any other model
- **Variable inclusion frequency (VIF):** how likely is selection of a variable?
- **Pairwise inclusion frequency (PIF):** how likely is selection of a pair of variables?  
(Sauerbrei & Schumacher, 1992)
- **Relative bias conditional on selection:** % bias in coefficient if variable is selected  
(RCB)
- **Root-mean-squared-difference ratio:** inflation ( $>1$ ) or deflation ( $<1$ ) of MS by selection (compared to full model)  
(RMSDR)
- **Stability paths**  
useful to assess dependence of inclusion on inclusion threshold  
(Meinshausen & Bühlmann, 2010)



## Resampling-based indicators of model (in)stability

- From Wallisch et al, StatMed 2021:

**TABLE 1** Estimands describing the uncertainty of model estimation incurred by variable selection, their approximation by simulation, and their estimation by resampling

Estimand	Definition	Approximation by simulation	Resampling-based estimator
VIF <sub>j</sub>	$E[I(\hat{\beta}_j \neq 0)]$	$\sum_{q=1}^Q I(\hat{\beta}_j^q \neq 0)/Q$	$\sum_{b=1}^B I(\hat{\beta}_j^b \neq 0)/B$
MSF( <i>J</i> )	$E[\prod_{j \in J} I(\hat{\beta}_j \neq 0) \cdot \prod_{j \in J'} I(\hat{\beta}_j = 0)]$	$\sum_{q=1}^Q \prod_{j \in J} I(\hat{\beta}_j^q \neq 0) \cdot \prod_{j \in J'} I(\hat{\beta}_j^q = 0)/Q$	$\sum_{b=1}^B \prod_{j \in J} I(\hat{\beta}_j^b \neq 0) \cdot \prod_{j \in J'} I(\hat{\beta}_j^b = 0)/B$
RCB <sub>j</sub>	$\left( \frac{E(\hat{\beta}_j)}{\beta_j \cdot VIF_j} - 1 \right)$	$\left( \frac{\sum_{q=1}^Q \hat{\beta}_j^q}{\beta_j \cdot VIF_j \cdot Q} - 1 \right)$	$\left( \frac{\sum_{b=1}^B \hat{\beta}_j^b}{\beta_j \cdot VIF_j \cdot B} - 1 \right)$
RMSDR <sub>j</sub>	$\sqrt{\frac{E(\hat{\beta}_j - \beta_j)^2}{E(\hat{\beta}_j - \beta_j)^2}}$	$\sqrt{\frac{\sum_{q=1}^Q (\hat{\beta}_j^q - \beta_j)^2/Q}{\sum_{q=1}^Q (\hat{\beta}_j^q - \beta_j)^2/Q}}$	$\sqrt{\frac{\sum_{b=1}^B (\hat{\beta}_j^b - \beta_j)^2/B}{\hat{\sigma}_j^2}}$

Note: Superscripts *q* or *b* indicate estimates obtained in the *q*th simulated dataset or the *b*th resample, respectively. Estimates  $\hat{\beta}_j$  and  $\hat{\beta}_j^q$  are set to 0 if they are not selected by the variable selection algorithm in the corresponding model. For the sets *J* and *J'* of indices,  $J \cup J' = \{1, \dots, k\}$ , and  $J \cap J' = \{\}$ , *I*(·) is the indicator function, that is, it is 1 if the expression (·) is true and 0 otherwise.

Abbreviations: MSF, model selection frequency; RCB, relative conditional bias; RMSDR, root mean squared difference ratio; VIF, variable inclusion frequency.

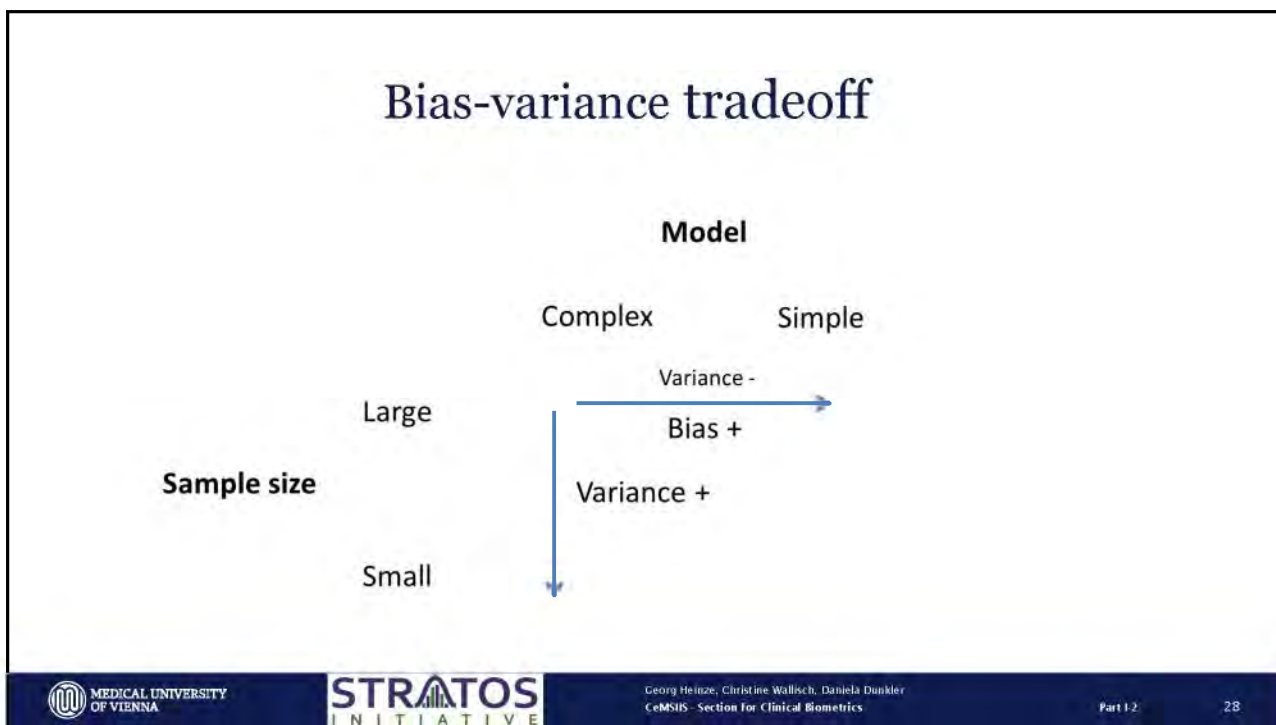
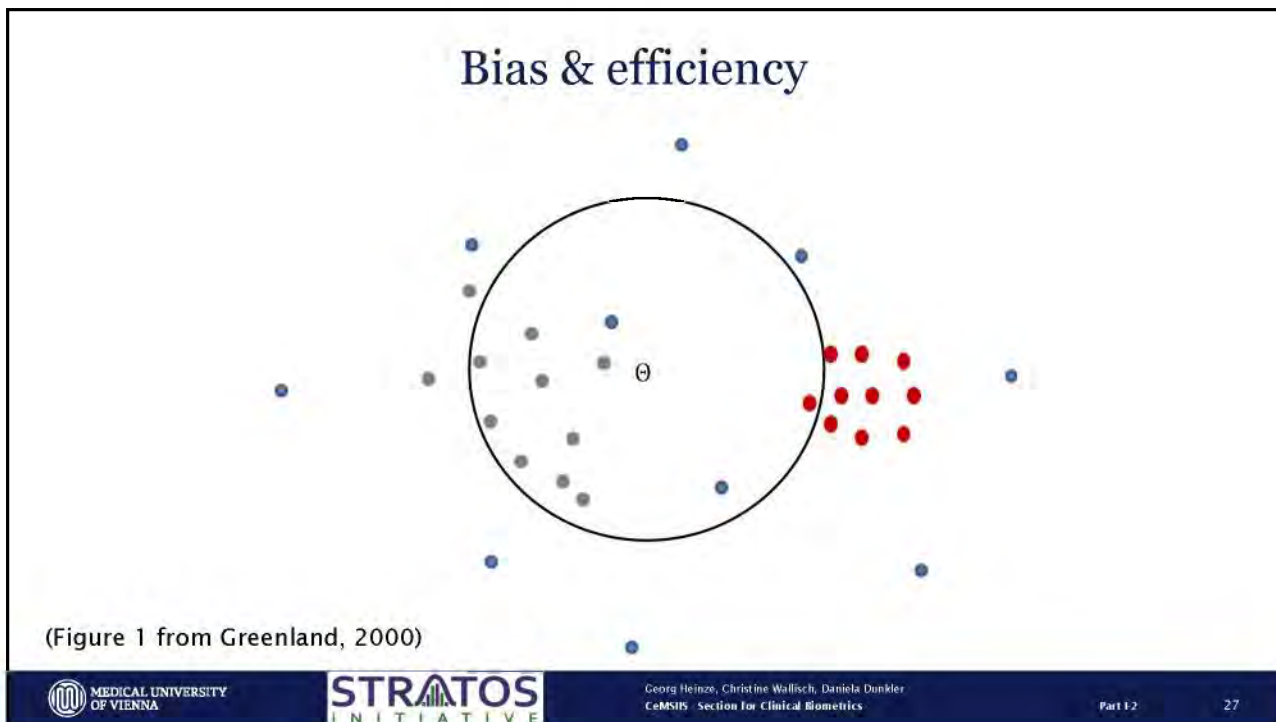
## Resampling methods

### Bootstrap

- Draw *B* samples with replacement from original data set.
- Perform model selection on each sample.
- Use for RMSDR, RCB (Wallisch, 2021)

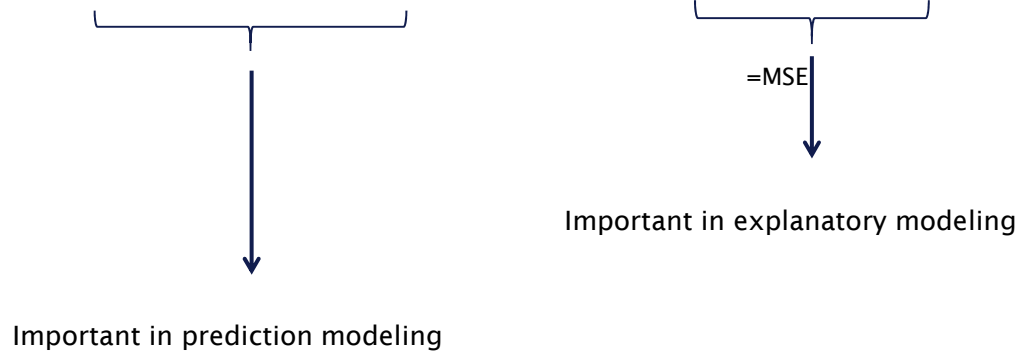
### Subsampling

- Draw *B* samples of size *M* < *N* without replacement.
- Perform model selection on each sample.
- *M* = *N*/2 yields distributions of regression coefficients similar to the bootstrap
- Use for VIF, MSF (Wallisch, 2021)



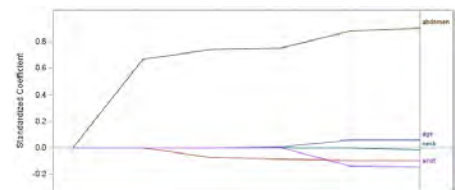
## To explain or to predict?

$$\text{Expected prediction error} = \text{Irreducible error} + \text{Bias}^2 + \text{Variance}$$



## Penalized likelihood: regularized regression

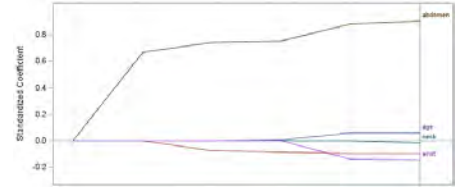
- LASSO: minimize  $\sum_i (y_i - \hat{y})^2 + \lambda \sum |\beta_j|$
- Imposes a penalty on the regression coefficients.



- Many variants/new methods (see Desboulets, Econometrics 2018 for a review)
- Prerequisite: adequate standardization of variables.
- What we obtain
  - A prediction formula with less error than ordinary least squares,
  - Shrinkage (to the mean),
  - Variable selection.

## Penalized likelihood: regularized regression

- LASSO: minimize  $\sum_i (y_i - \hat{y})^2 + \lambda \sum |\beta_j|$
- Imposes a penalty on the regression coefficients.

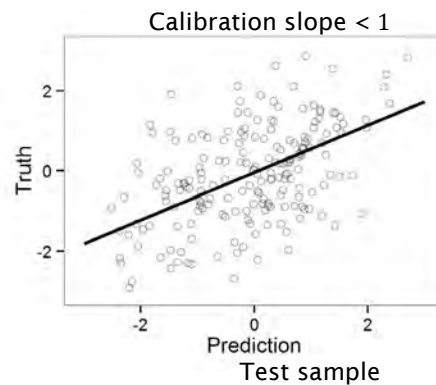
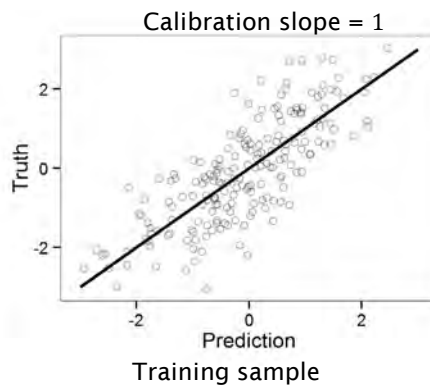


- What we do not obtain
  - Unbiased regression coefficients,
  - Confidence intervals:
    - Sampling distribution can be assessed with bootstrap, but because of the bias, does not give valid confidence intervals.
    - Post-selection inference: still in its infancy (Kammer, arXiv 2021)

## Shrinkage

### The phenomenon

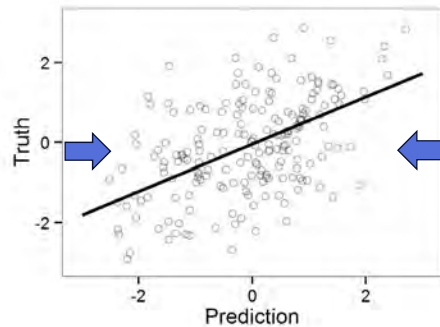
- Observed values in new samples are closer to overall mean than predicted values.



## Shrinkage

### The method(s)

- Anticipate shrinkage (of calibration slope) by cross-validation
- ‘Shrink’ regression coefficients such that a calibration slope of 1 would be expected.



Calibration slope < 1

## Shrinkage methods

- **Post-estimation shrinkage factor estimation**
  - Verweij & Van Houwelingen 1993: global shrinkage factor  $c$  ( $c < 0.8 \rightarrow$  poor model)
  - Sauerbrei, 1999: parameterwise shrinkage factors
  - Dunkler, 2016: joint shrinkage factors, R package `shrink`
- **Regularized regression**
  - Ridge regression: L2 penalty on regression coefficients
  - Lasso: L1 penalty (Tibshirani, 1996 & 2011)
  - Elastic net: L2 and L1 penalty
- **Better prespecify than cross-validate penalty strength?**
  - Greenland, StatMed 1997 (Empirical Bayes vs. Semi-Bayes)
  - Van Calster et al, SMMR 2020
  - Riley et al, JCE 2021
  - Sinkovec et al, upcoming in BMC MedResMeth 2021 (‘To tune or not to tune’)

## Shrinkage

- Consequences of shrinkage:
  - Controlling variance, not bias.
  - Inference about effects after shrinkage?
- Selection = extreme shrinkage!
  - “If it’s close to 0, set it to 0.”
- Not to be confused with bias correction!
  - It does not aim at unbiased regression coefficients!
- It may decrease the overall MSE, but can lead to higher local MSE (see later)



## Addressing confounding in descriptive models (?)

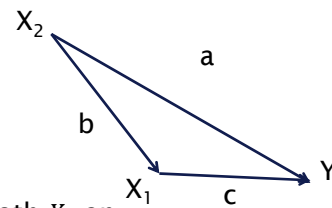
- Explanatory models: we have to consider confounding (‘we must’)
- Descriptive models: we choose our confounders to adjust our estimates (‘we want’)

### Directed acyclic graph (DAG)

- = A graph with one-way edges containing no cycles describing causal relationships.

### Confounding

- Effect of  $X_1$  on  $Y$  is confounded by  $X_2$ , if  $X_2$  is effect of both  $X_1$  and  $Y$ .
- →  $X_2$  must be considered to regain causal interpretation of effect of  $X_1$  on  $Y$ . (Pearl, 1995)



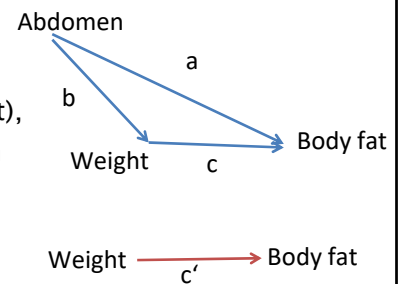


## Change-in-estimate criterion

- In epidemiologic studies, it is often not clear whether adjustment for a variable  $X_2$  is necessary or not.
- In descriptive model, we can decide → impacts interpretation

### Change-in-estimate criterion

- If  $X_2$  (abdomen circumference) is a confounder ( $a$  and  $b$  exist), then its removal will change our assessment of arrow  $c$  from weight to body fat.
- So we could remove 'abdomen' and see what happens to  $c$ : CIE =  $c' - c$ .



## Change-in-estimate criterion

- $M_1: \beta_0 + \beta_1 X_1 + \beta_2 X_2$
- $M_2: \theta_0 + \theta_1 X_1$
- Change in estimate criterion: leave  $X_2$  in the model if  $\beta_1 - \theta_1 \neq 0$ , often proxied by

$$\text{abs}(\hat{\theta}_1 - \hat{\beta}_1) / \hat{\beta}_1 > 0.10$$

- This leads to inconsistent variable selection (Maldonado & Greenland, 1993)
- To get a consistent estimator, we could test for  $\beta_1 \neq \theta_1$  (collapsibility of the two models).

(see also Lee, 2014)

## Significance of change-in-estimate

- Tests for collapsibility by bootstrapping or approximation
- Dunkler et al (2014) approximate the change-in-estimate and derive a simple test for  $\beta_1 - \theta_1 = 0$ .

We showed:

- Elimination of a ,significant' variable  $X_2$  from a model leads to a significant change  $\hat{\beta}_1 - \hat{\theta}_1$ .
- Elimination of a 'non-significant' variable  $X_2$  from a model leads to a non-significant change  $\hat{\beta}_1 - \hat{\theta}_1$ .
- → Test of collapsibility = Test of omitted variable.
- Compare to Greenland, Daniel & Pearce (IJE 2016): 'change in MSE'

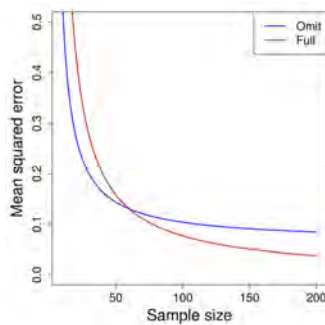
## Motivation for omission: to reduce MSE?

- Omission of  $X_2$  successful (in terms of MSE of  $\beta_1$ ) if:

$$\text{Bias}_{\text{omit}}^2 < \text{Variance}_{\text{full}} - \text{Variance}_{\text{omit}}$$

Independent of  $N$

Inversely proportional to  $N$



Success of ,always omit' depends on sample size

Luijken et al, *submitted*:

A comparison of full model specification and backward elimination of potential confounders when estimating marginal and conditional causal effects on binary outcomes from observational data

## Motivation for selection (adaptive omission): omit weak effect to reduce MSE

- Simulation with  $N = 50$

True  $\beta_1 = 1.5, \beta_2 = 0.3$

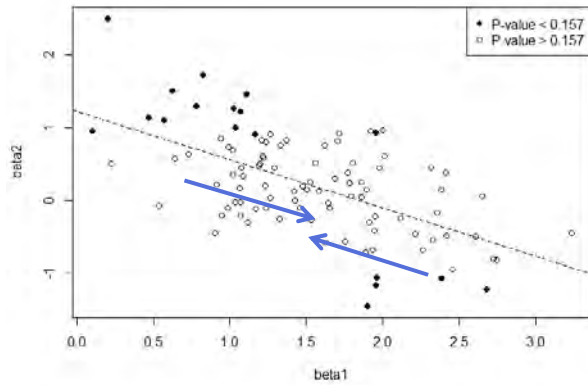
A weak  $\beta_2$ :  
Setting it to 0 will more often push  $\hat{\beta}_1$   
towards its true value than away from it.  
Shrinkage effect on  $\hat{\beta}_1$ !

$$\text{RMSE}(\hat{\beta}_{1,FULL}) = 0.67$$

$$\text{RMSE}(\hat{\beta}_{1,BE}) = 0.65$$

$$\text{Bias}(\hat{\beta}_{1,FULL}) = -0.03$$

$$\text{Bias}(\hat{\beta}_{1,BE}) = +0.03$$



→ 'Selection is good.'

## Motivation for selection (adaptive omission): poor results if applied on strong predictor

- Simulation with  $N = 50$

True  $\beta_1 = 1.5, \beta_2 = 1.5$

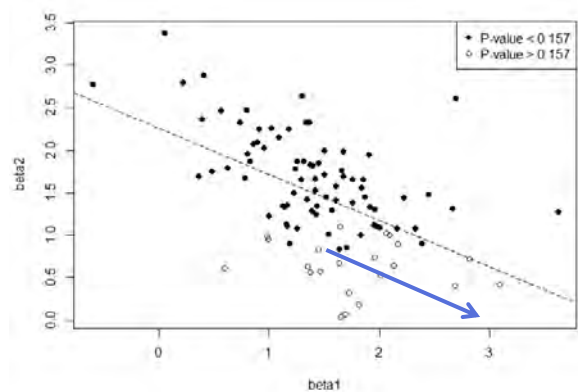
A strong  $\beta_2$ :  
Setting it to 0 will always push  $\hat{\beta}_1$   
from its true value.

$$\text{RMSE}(\hat{\beta}_{1,FULL}) = 0.68$$

$$\text{RMSE}(\hat{\beta}_{1,BE}) = 0.67$$

$$\text{Bias}(\hat{\beta}_{1,FULL}) = -0.03$$

$$\text{Bias}(\hat{\beta}_{1,BE}) = +0.33$$

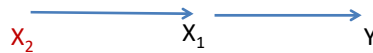


→ 'Selection is bad.'

## Prior knowledge

We should have known the likely role of  $X_2$  in advance:

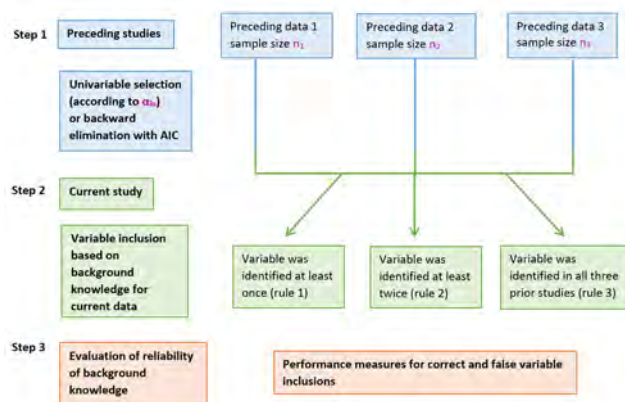
- If it is considered strongly associated with  $Y$ , never consider deletion from the model!
- If it is considered weakly associated with  $Y$ , selection can improve performance. → smaller variance (Shmueli, 2010)
- If it is considered not associated with  $Y$ , it better should not have been used upfront ('instrumental variable').

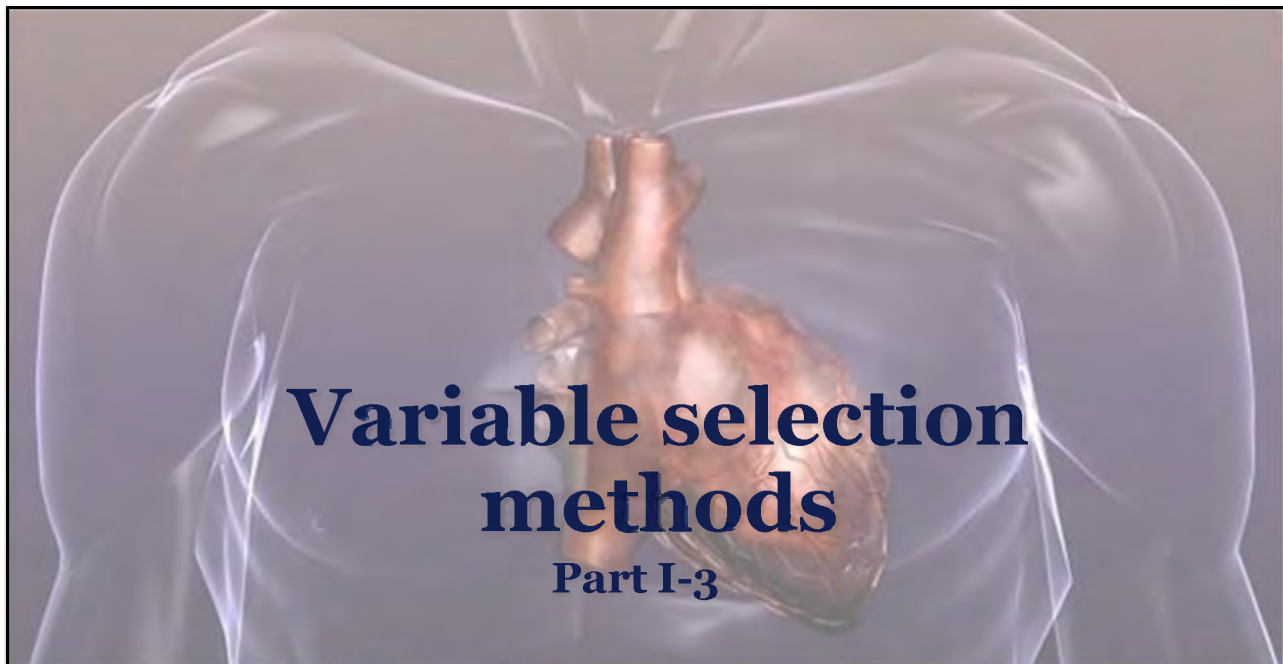


## Dangers of using prior ,knowledge‘

- Beware of prior knowledge from poorly conducted preceding studies:
- These studies may have used inappropriate selection methods


- cf. Hafermann et al: Statistical Model Building: Background ,Knowledge‘ based on inappropriate preselection causes misspecification (upcoming in BMC Med Res Meth 2021)






# Variable selection methods

## Part I-3

 MEDICAL UNIVERSITY OF VIENNA

 STRATOS INITIATIVE

Georg Heinze, Christine Wallisch, Daniela Dunkler  
CeMSIS - Section for Clinical Biometrics

Part I-3

## Aims

- Distinguish expertise-based preselection from data-driven selection
- Understand motivation for data-driven selection as connected to the aim of modeling
- Different recommendations in the literature may be explained by differences in the set of assumptions on sample size, number of candidate variables, modeling aim, level of expertise, ... .

 MEDICAL UNIVERSITY OF VIENNA

 STRATOS INITIATIVE

Georg Heinze, Christine Wallisch, Daniela Dunkler  
CeMSIS - Section for Clinical Biometrics

Part I-3

2

## Basic algorithms

- 'Full' model specification
- Univariable filtering
- Best subset selection
- Forward selection
- Backward elimination
- Full model approximation
- Change-in-estimate: Purposeful variable selection and augmented backward selection
- Information-theoretic approach
- Directed acyclic graph (DAG)-based selection

## Our basic approach: from the 'Full' to a meaningful 'Global' model

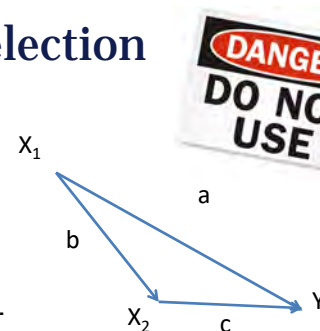
- 'Full model' means: do not perform any data-driven variable selection, take what you have.
- 'Global model' means the largest meaningful model.
- Be sure about purpose of analysis.
- Use domain expertise to remove some variables upfront.
- Select, for each variable, a desired level of non-linearity (including spline transformations).
- Select some biologically plausible interactions.

## Univariable filtering

- Still by far the most often applied variable selection method in medical literature!
- Select a significance level  $\alpha$  (e.g.,  $\alpha=0.20$  or  $\alpha=0.157$ )
- Perform  $K$  univariable models.
- Use all variables in multivariable model with univariable  $p$ -value  $< \alpha$ .
- Sometimes accompanied by subsequent backward elimination.
- Popular: ~25% of COVID-19 prediction models employed univariable filtering (Wynants, BMJ 2020)

## Pros and cons of univariate selection

- Easy. (You can do that with any software.)
- Retractable.
- Problematic (Sun et al 1996):
  - Selection is on the total effect of  $X_1$  on  $Y$  ( $a + bc$ )
  - In multivariable context, only direct effect ( $a$ ) is of interest.
  - Based on most likely misspecified models: total effect is estimated while ignoring adjustment for any other variables.



a	b	c	Consequence
Pos.	Pos.	Neg.	$X_1$ falsely not selected (if $a = -bc$ )
0	Pos./Neg.	Pos./Neg.	$X_1$ falsely selected.
Pos./neg	0	Pos./neg	$X_1$ correctly selected (only if $b = 0$ or $c = 0$ ).

## Best subset selection

- Perform all  $2^K$  regressions.
- Select the model that has the lowest AIC.

### Modification:

- Pre-specify a small number (4 – 20) of plausible models.
- Select those that have  $AIC < AIC_{\min} + 2$ .
- Perform multi-model inference on the selected models.

### In practice:

- Approximated by stepwise approaches!

(Burnham & Anderson, 2002)

## Forward selection

- Select a significance level  $\alpha_1$ .
- Repeat:
  - While the most significant currently excluded term has  $p < \alpha_1$ , add it and re-estimate.

Software:  
SAS/PROC GLMSELECT  
R `step()`

### Variant: Stepwise forward

- If least significant included term has  $p \geq \alpha_2$ , remove it and re-estimate.
- Problem:
  - Starts with grossly misspecified models



## Backward elimination

- Select a significance level  $\alpha_2$ .
- Estimate full model.
- Repeat:
  - While least significant term has  $p \geq \alpha_2$ , remove it and re-estimate.

### Variant: Stepwise backward

- If most significant excluded term has  $p < \alpha_1$ , add it and re-estimate.

### Software:

R `mfp:mfp()`

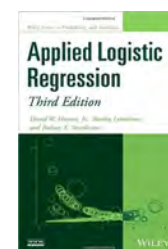
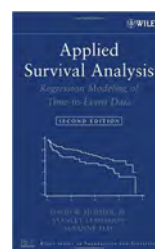
- Should start with plausible ,global' model

## Full model approximation

- Select a level of approximation  $R_{red}^2$  from  $[0, 1]$  (typically  $R_{red}^2=0.95$ ).
- Fit full model.
- Compute linear predictors  $\eta$  from full model.
- Fit model with  $\eta$  as dependent variable ( $R^2 = 1$ )
- Repeat:
  - Remove variables sequentially as long  $R^2 \geq R_{red}^2$
- Suggested in Harrell (2015)
- Exemplified in Cowling et al (JCE 2020).

## Purposeful selection

- Proposed by Hosmer and Lemeshow in their books on applied logistic regression and applied survival analysis.
- Starts with univariate screening.
- Then performs backward elimination, but leaves variables in the model if omission would cause a large (proportional) change-in-estimate in other variables.
- Additional forward steps.
- A bit outdated.



(Hosmer & Lemeshow, 2011 & 2013)

## Augmented backward elimination

- Proposed by Dunkler et al, 2014.
- Re-investigated the change-in-estimate criterion and proposed a standardized version and a short-cut approximation to it.
- Based on backward elimination with level  $\alpha_2$ .
- Leaves variable in a model if maximum of standardized changes-in-estimate greater than  $\tau$ .
- Simulation study showed that results and performance are always close to the full model, but fewer variables are selected.

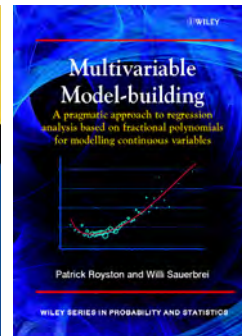
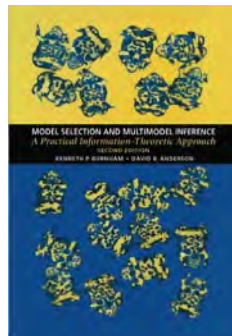
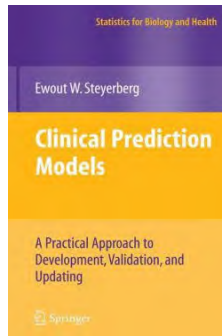
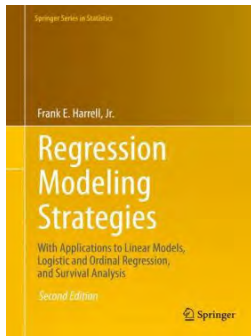
Software: SAS macro %ABE, R package **abe**

## Opinions on variable selection

- for models with focus on prediction and description.



Variable selection



(Harrell, 2001; Steyerberg, 2009; Burnham & Anderson, 2002, Royston & Sauerbrei, 2008)

## Harrell's recommendations

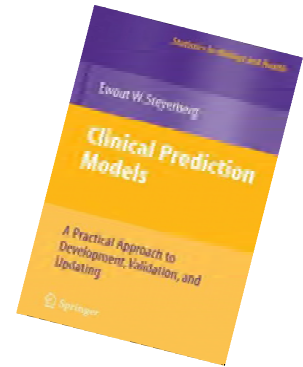
- Focus on prediction models.
- 'Effects cannot be assumed to be exactly 0.'
- 'Selection invalidates confidence intervals and p-values.'
- Specify a full model, including meaningful interactions and non-linear effects.
- Perform global tests for interactions or non-linear effects.
- At most: do a mild backward selection at  $\alpha_2 = 0.50$ .
- Model simplification using cross-validated predicted values as outcome.



(see also Harrell, 1996)

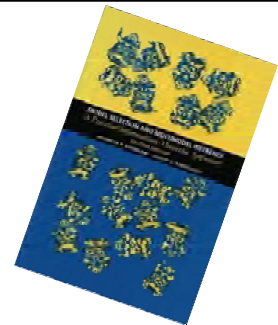
## Steyerberg's recommendations

- Focus on prediction models.
- False inclusion is better than false exclusion of variables.
- Stepwise methods may lead to
  - Instability of selection,
  - Biased estimation of coefficients,
  - Misspecification of variability (exaggerated  $p$ -values),
  - Predictions of worse quality than from a full model.



## Burnham-Anderson's recommendations

- Strong focus on descriptive (,explanatory\*') models.
- Select a set of models that are biologically plausible.
- These are subset models of a global model.
- Apply information-theoretic approach.
- Compute AIC weights or bootstrap weights.
 
$$\Delta_i = AIC_i - AIC_{\min}$$
 Akaike weight:  $w_i = \frac{\exp(-\Delta_i/2)}{\sum_r \exp(-\Delta_r/2)}$
- Perform multi-model inference (problem: no variable selection!).



Model	$\Delta_i$	$\mathcal{L}(M_i x)$	$w_i$
1	0	1	0.431
2	1.2	0.5488	0.237
3	1.9	0.3867	0.167
4	3.5	0.1738	0.075
5	4.1	0.1287	0.056
6	5.8	0.0550	0.024
7	7.3	0.0260	0.010

90% confidence set ←

\* Burnham&Anderson's definition of ,explanatory' differs from Shmueli's

## Model averaging

- $\bar{\beta}_j = \frac{\sum_r \hat{\beta}_{j,r} I_{r,j} w_{j,r}}{w^+(j)}$        $I_{r,j}$  ... inclusion of  $\beta_j$  in model  $r$   
 $w^+(j)$  ... sum of weights of models including  $\beta_j$

- $\widehat{\text{var}}(\bar{\beta}_j) = \left[ \underbrace{\sum_r w_r}_{\text{weight}} \underbrace{\widehat{\text{var}}(\hat{\beta}_{j,r} | M_r)}_{\text{within-model variance}} + \underbrace{(\hat{\beta}_{j,r} - \bar{\beta}_j)^2}_{\text{between-model variance}} \right]^2$

(Buckland, 1997)

## Burnham-Anderson's recommendations

### For descriptive (explanatory\*) model

- If there is a dominating model with  $w_i > 0.9$ , just report this one unconditionally.
- Otherwise, report the best performing model, with unconditional variance based on model-averaged inference on the models of the 90% confidence set.

### For prediction model

- Perform model-averaged inference (averaged point estimate and variance).

Bootstrap model frequencies can replace the Akaike weights.

Relative importance of a variable  $X_j$ :  $w^+(j) = \sum_r w_j I_{j,r}$

\* Burnham&Anderson's definition of ,explanatory' differs from Shmueli

## Royston-Sauerbrei 's recommendations

- Focus on descriptive models.
- Initial working set of variables.
- Coding matters.
- Backward elimination with additional forward steps.
- Function selection. (not covered here)
- 'If you have a large enough sample, you can use selection methods.'
- They propose backward elimination.
- Select  $\alpha_2$  according to needs; larger value means larger model.
- Emphasize importance of investigation of model stability → by means of resampling.



## Coding

- One interesting aspect (out of many) in the Royston-Sauerbrei (2008) book is their discussion of appropriate coding of categorical variables:
- Nominal variables: choose an appropriate reference.
  - Frequent, standard group, etc.
  - Variable selection on dummies – collapse rare groups with reference
- Ordinal variables: advantages of ordinal coding
  - Variable selection can then collapse adjacent groups with similar outcome

Reference coding:

Level	Dummy 1	Dummy 2
0	0	0
1	1	0
2	0	1
etc.		

Ordinal coding:

Level	Dummy 1	Dummy 2
0	0	0
1	1	0
2	1	1
etc.		

## Differences (and similarities) in prediction, explanatory and descriptive modeling

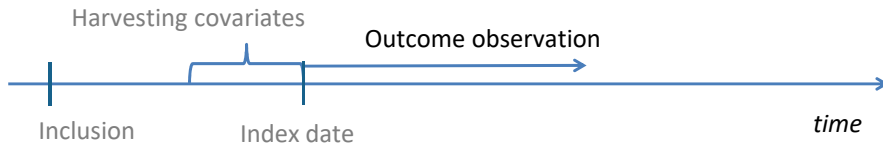
- All are using maximum likelihood → prediction as vehicle to find estimates.
- While prediction focusses on  $\hat{Y}$  and the minimization of prediction error, causal modeling focusses on causal contrasts such as  $\hat{\beta}$  or, more generally,  $E(Y|X = 1) - E(Y|X = 0)$
- In prediction, important prerequisites for selecting variables are:
  - Chronology (do not use future values!, e.g. time-dependent variables in survival analysis),
  - Availability at time of prediction.
- In explanatory modeling, it is confounder control. +/- error minimization → DAG methodology
- In descriptive models, it is mostly simple interpretation.

## „Expertise“-driven preselection

Modeling aim	Expertise preselection of independent variables	What data-driven selection may add
Prediction	Availability, chronology, costs, assumed associations with Y	Remove weak candidate predictors to decrease MSE
Explanation	Identify causal contrast of interest by appropriate confounder control	Remove „instruments“ to decrease MSE
Description	What are the variables I want to consider?	Reduce model size (parsimony), Remove weak predictors to decrease MSE

## Preselection for (prognostic) prediction models

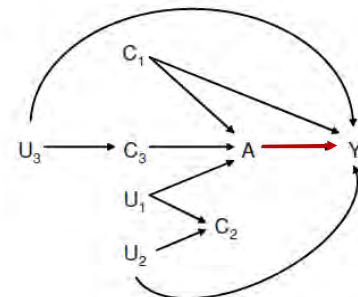
- Chronology:



- Don't use information from the future for prediction! (see e.g. Wynants, BMJ 2020)  
(This is one of the most often violated conditions in practice!)
- $X$  must be available also in prediction situation.

## Using causal DAGs to identify confounders

- Pearl (1995) described causal relationships by DAGs.
- We are interested in the effect of  $A$  on  $Y$ .



Confounder adjustment should be made for:

- Confounders (parents of  $A$  and  $Y$ :  $C_1$ )
- Backdoor path blockers (they look like confounders:  $C_3$ )
- NOT for instruments ( $C_3$  if  $U_3$  were not there)
- NOT for colliders ( $C_2$ )

(BIAS)

(BIAS)

(VARIANCE)

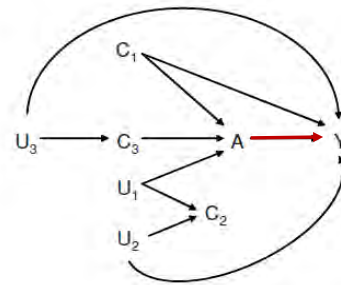
(BIAS)



## Implication of the DAG view on explanatory models

- **This implies that there cannot be a single model ,explaining  $Y'$ ,** but the choice of model depends on what we want to estimate:  
e.g., the causal effect of  $A$  on  $Y$ .

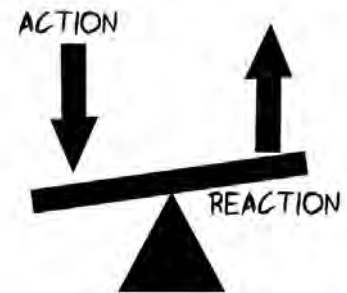
- If we were interested in the effect of  $C_1$  on  $Y$ , we would not adjust for  $A$  (and probably not for any other variable).



## Take home

- Distinguish expertise-based from data-driven variable selection
- Motivation for data-driven variable selection:
  - Make model more easily interpretable (description)
  - Make model more practically useful/communicable etc. (description, prediction)
  - Reduce noise (MSE) (description, prediction, explanation)
- Different recommendations stem from different contexts/scenarios:
  - Sample size
  - Purpose of modeling
  - Number of candidate variables
  - Assumptions/level of domain expertise
- Next, we will investigate consequences of algorithmic omission/inclusion  
→ are these theoretical considerations practically meaningful?

# Part II-1: Consequences of variable selection



<http://whatnextbook.com/wordpress/tag/decisions/>

## Questions

- How stable is variable selection?
- Does variable selection induce bias of  $\beta$ ?
- Does variable selection increase RMSE of  $\beta$ ?
- Does variable selection lead to biased or inaccurate predictions?
- How does background knowledge improve results?
  
- What is the role of sample size?

## Simulation - Aim

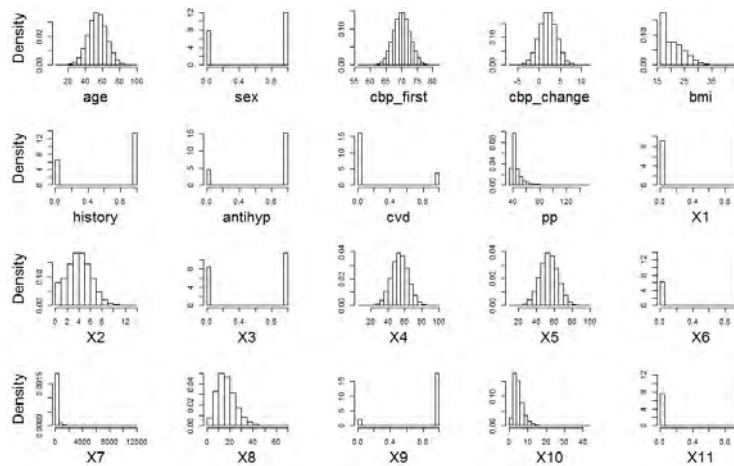
- Assessment of the impact of performing variable selection on regression coefficients and predictions obtained by the reduced model.

## Simulation – Data generating mechanism

- Scenario based on Sheppard, 2016 and Hafermann, 2021 mimicking a typical observational study in cardiology
- Outcome: difference in diastolic blood pressure between a measurement in a clinical environment and a measurement at home
- Predictors: age in years, sex, first reading of the clinical diastolic blood pressure, difference of the first and a follow-up reading of the clinical diastolic blood pressure, body mass index, hypertension, antihypertensive medication, pulse pressure, history of cardiovascular diseases
- 11 noise variables

## Simulation – Data generating mechanism

Univariate distributions of predictors



Part II-1

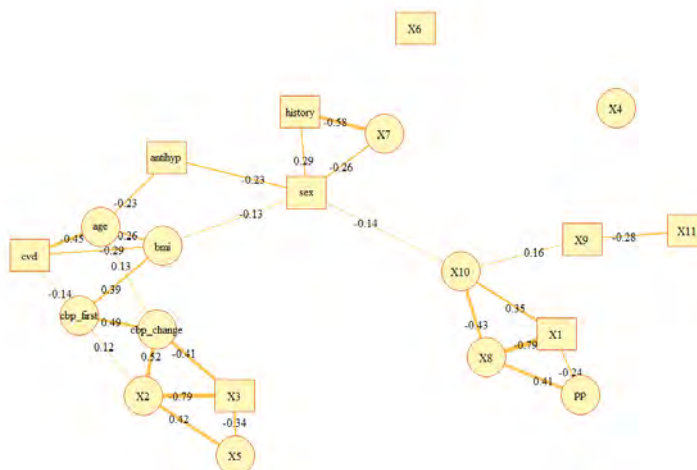
## Simulation – Data generating mechanism

Model to generate outcome variable

Variable	Unstandardized coefficients	Standardized coefficients	Partial R <sup>2</sup>
X <sub>sex</sub>	3.33	1.63	0.229
X <sub>cbp_first</sub>	-0.47	-1.27	0.103
X <sub>antihyp</sub>	2.37	1.00	0.091
X <sub>age</sub>	-0.08	-0.85	0.053
X <sub>cbp_change</sub>	0.31	0.63	0.029
X <sub>pp</sub>	-0.06	-0.52	0.029
X <sub>bmi</sub>	-0.07	-0.29	0.006
X <sub>cvd</sub>	-0.40	-0.16	0.002
Residual standard deviation	2.00		

Total R<sup>2</sup> 0.59

Correlations of predictors



Part II-1

## Simulation – Methods

- Univariable selection
  - with  $\alpha = 0.05$
  - with  $\alpha = 0.2$
- Forward stepwise selection with AIC (default value in step function in R)
- Backward elimination
  - with  $\alpha = 0.05$
  - with AIC
- Augmented backward elimination with  $\alpha = 0.2$  and  $\tau = 0.05$  (default values in abe function in R)
- Full model approximation with  $\omega = 0.95$
- Lasso with 10-fold CV lambda.min
- Relaxed Lasso with 10-fold CV lambda.min and subsequent ML estimates

## Simulation – Estimands & performance measures

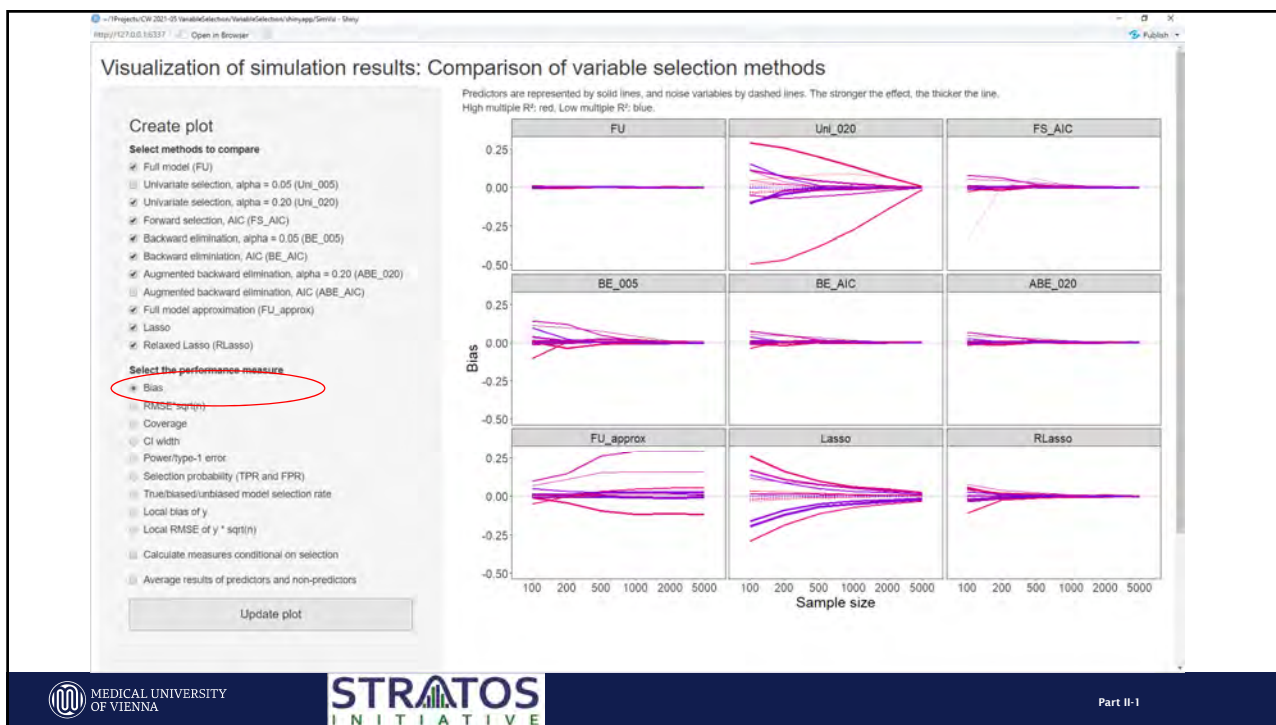
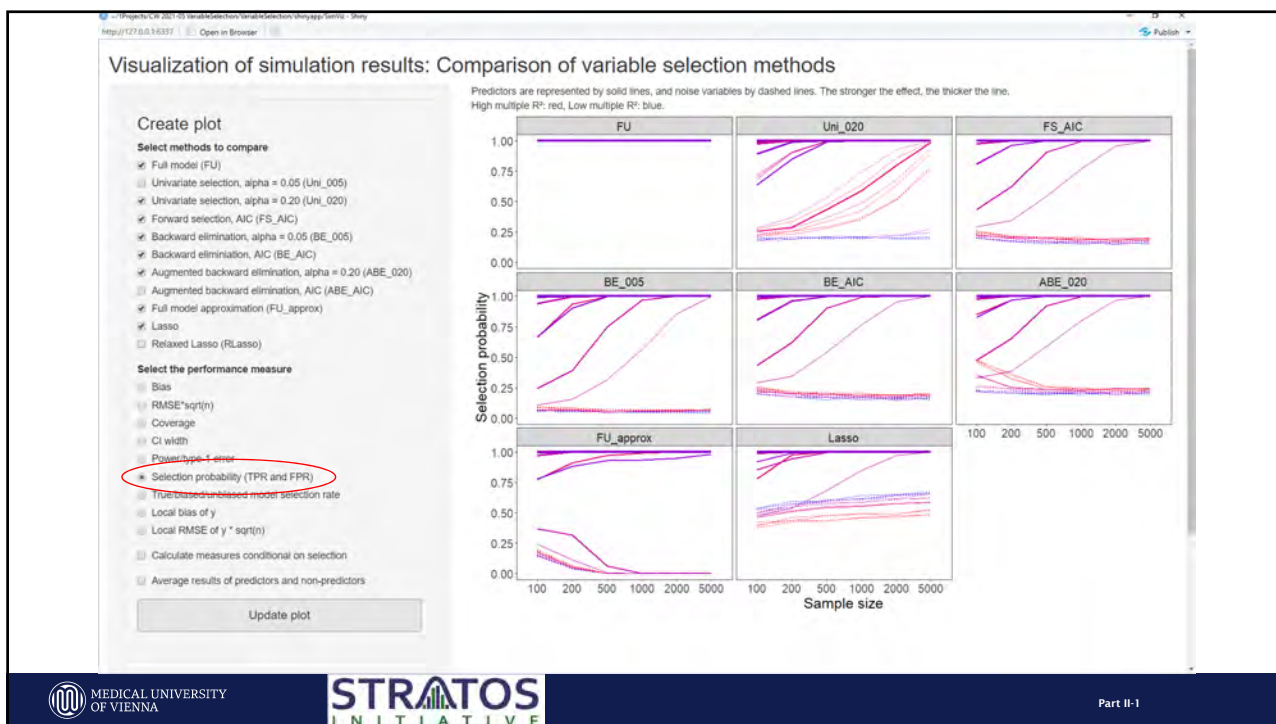
Estimand	Performance measure
$\beta_i$	Bias
	RMSE* $\sqrt{n}$
	Coverage of the 95% CI
	Width of the 95% CI
	Type 1 error / Power
	False positive rate / True positive rate
	Selection rate of the true/biased/unbiased model
$y$	Local bias
	Local prediction error

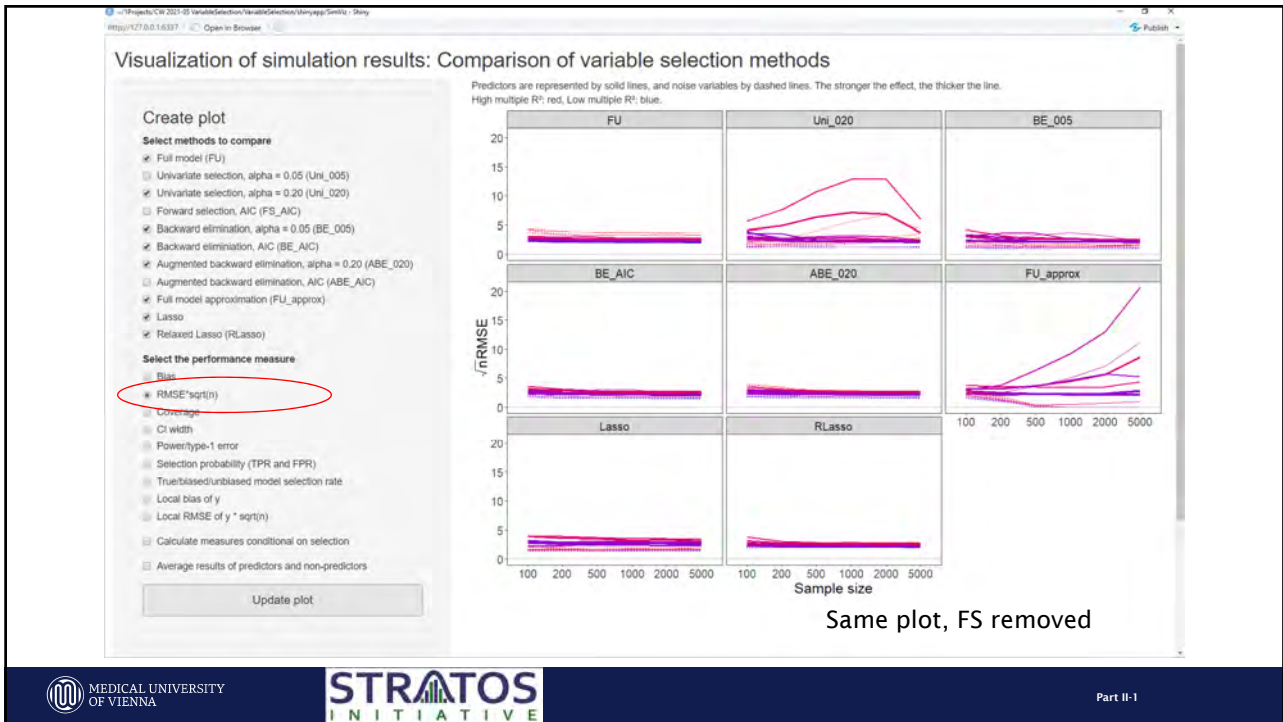
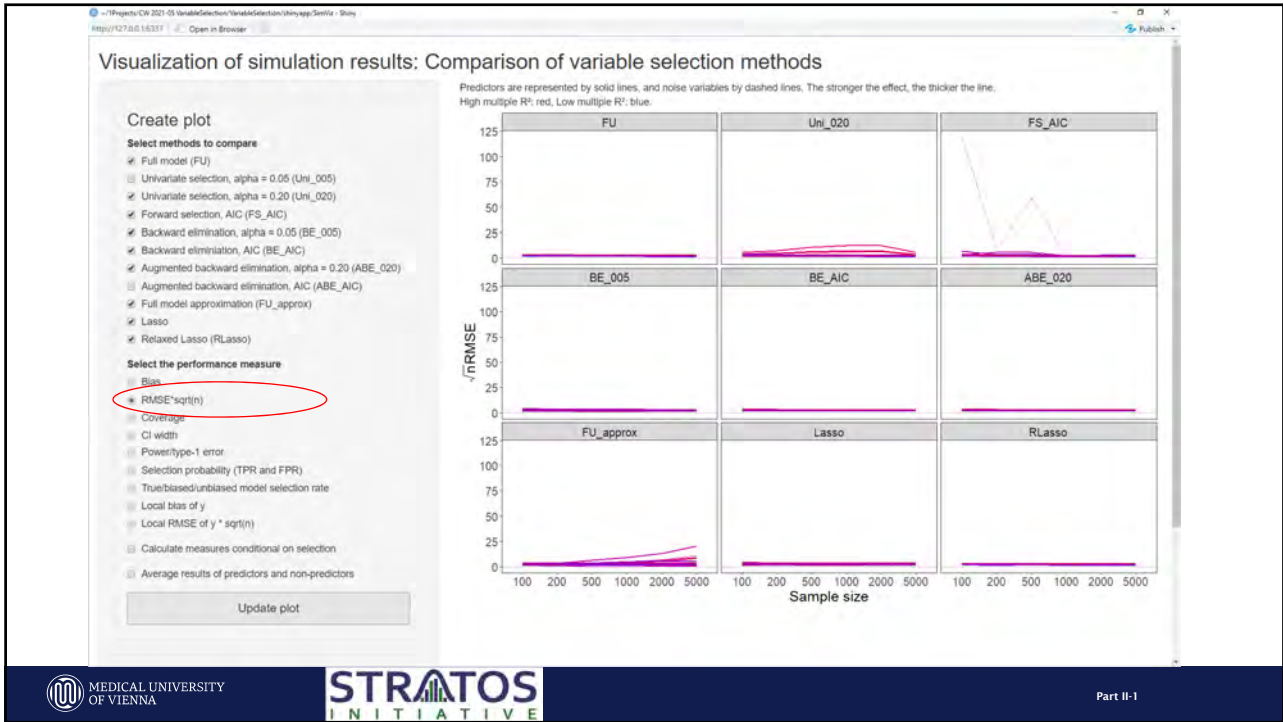
NB: we use RMSE\* $\sqrt{n}$  to equalize the effect of sample size

## A note of caution

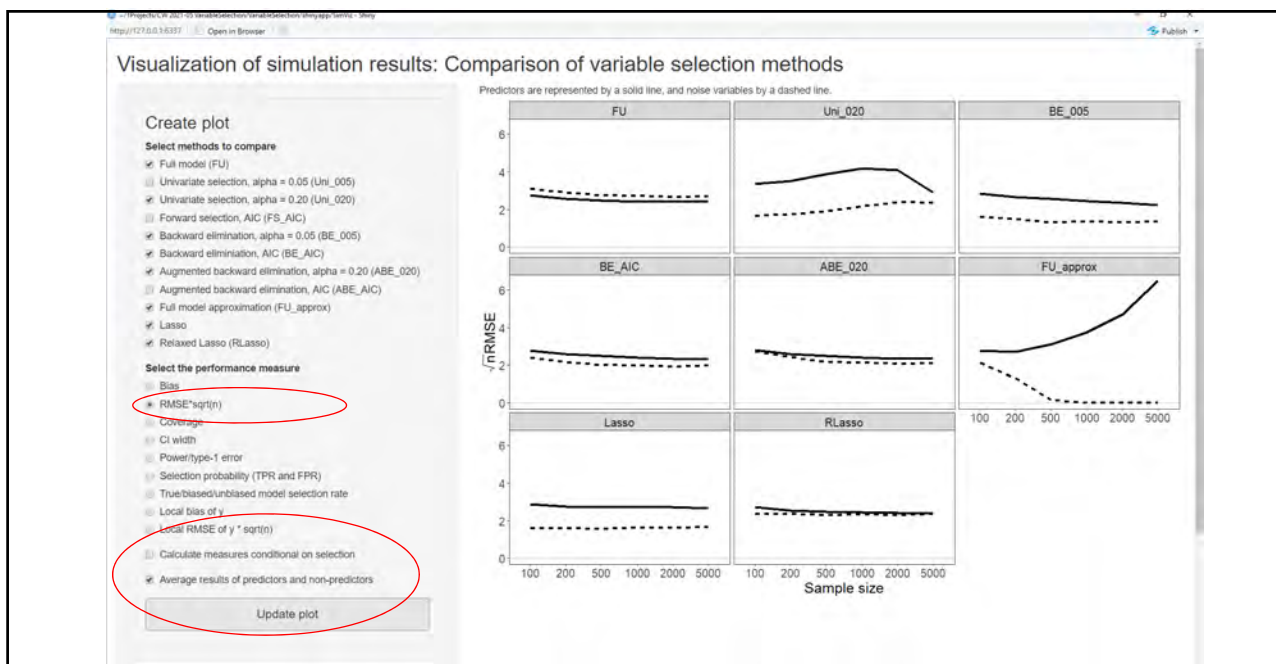
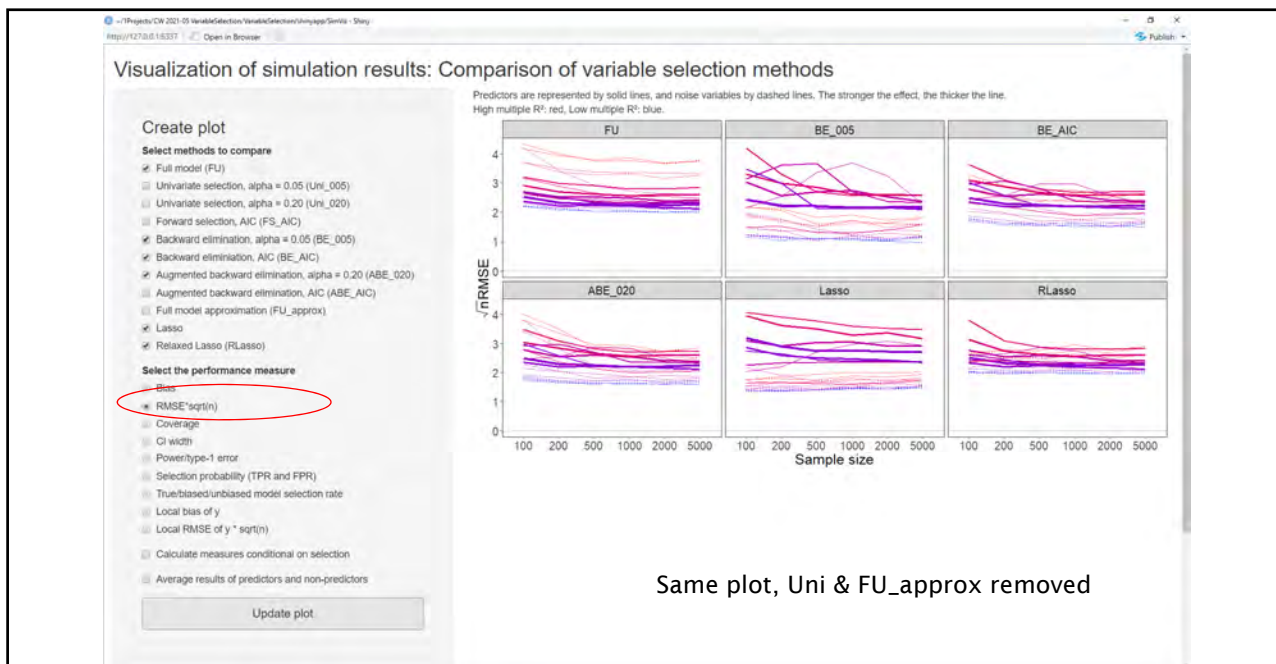
- We assume a ‘true model’ (data generating mechanism), even if we doubted its existence in Part I.
- We assume that a variable selection method may discover that ‘true model’.
- This way we can learn about the behavior of variable selection methods under known population properties.
- We can also evaluate ‘explanatory performance’ of the model (bias/RMSE of regression coefficients).
  
- Alternative way to compare methods:  
best cross-validated performance in complex data sets
  - Only possible for prediction performance
  - No general properties can be derived!

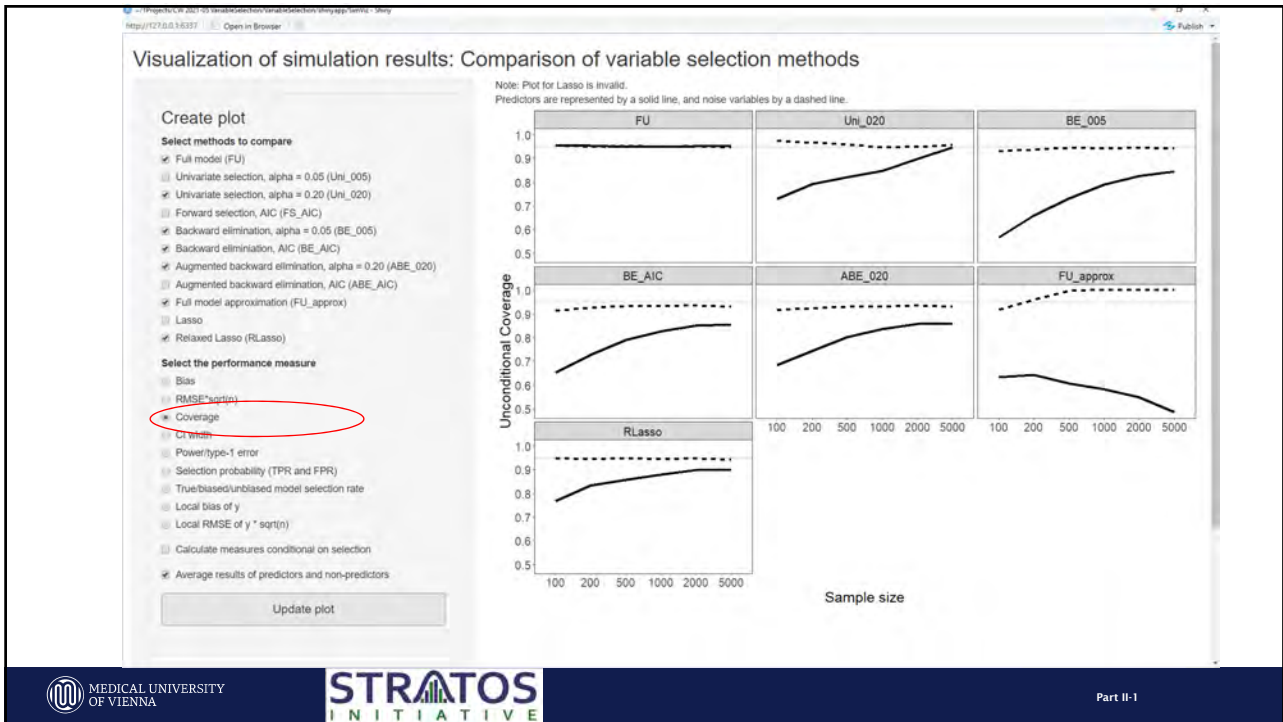
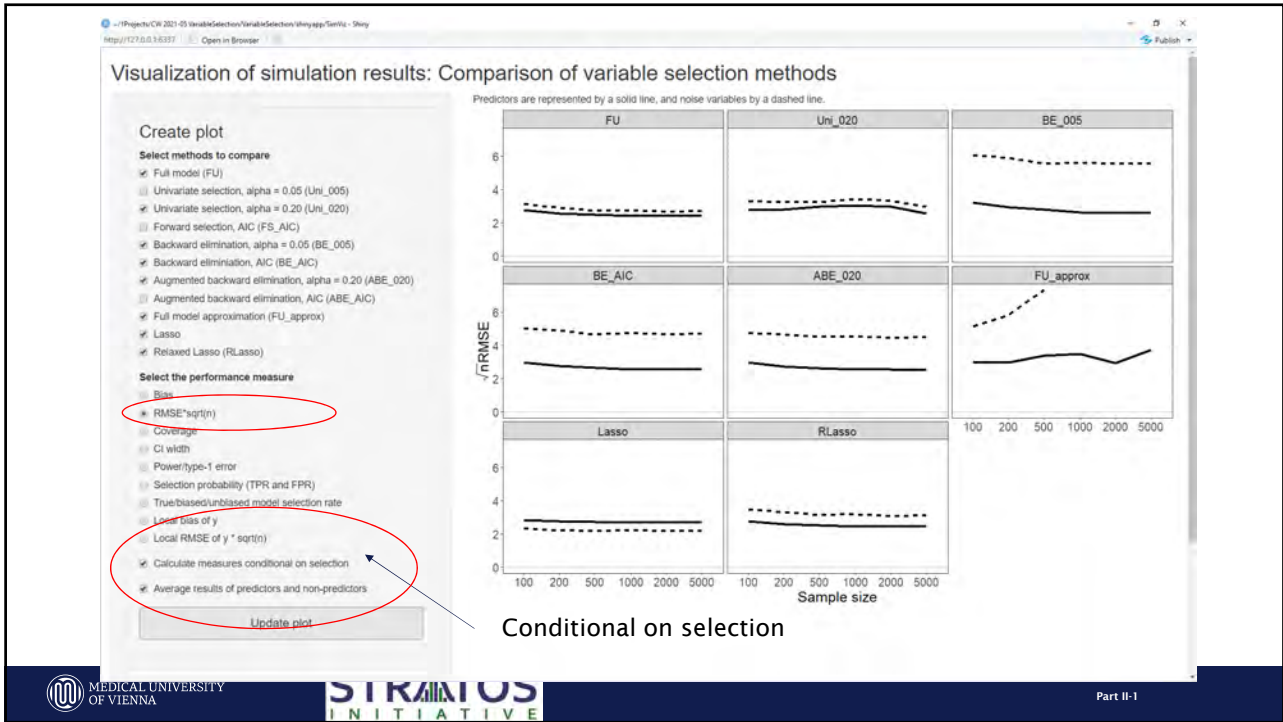
- *In the following we will show some screenshots of our shiny app with which results can be interactively browsed.*
- *During the workshop, we can extend these comparisons or show head-to-head comparisons between methods to better demonstrate relative performance of methods*

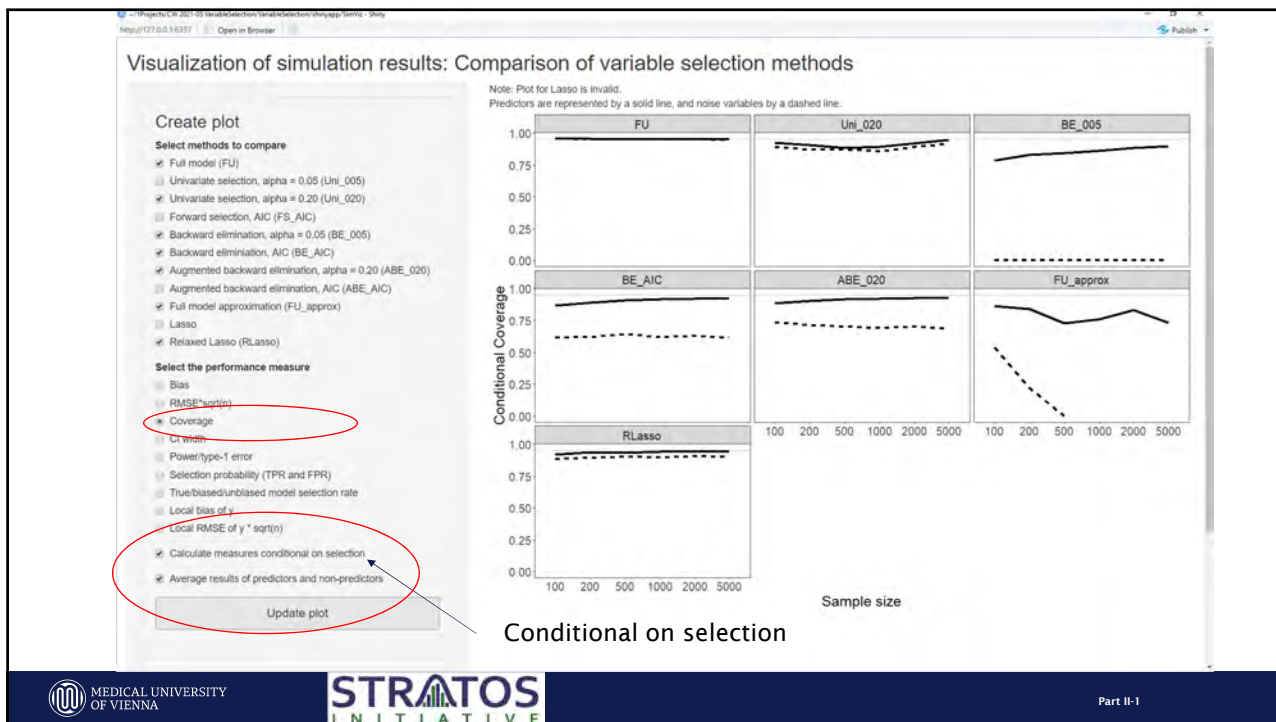




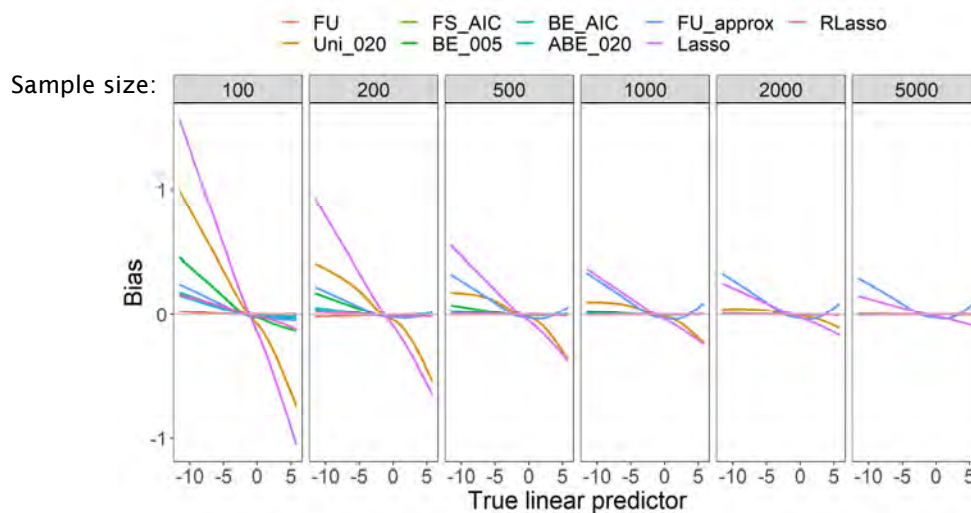




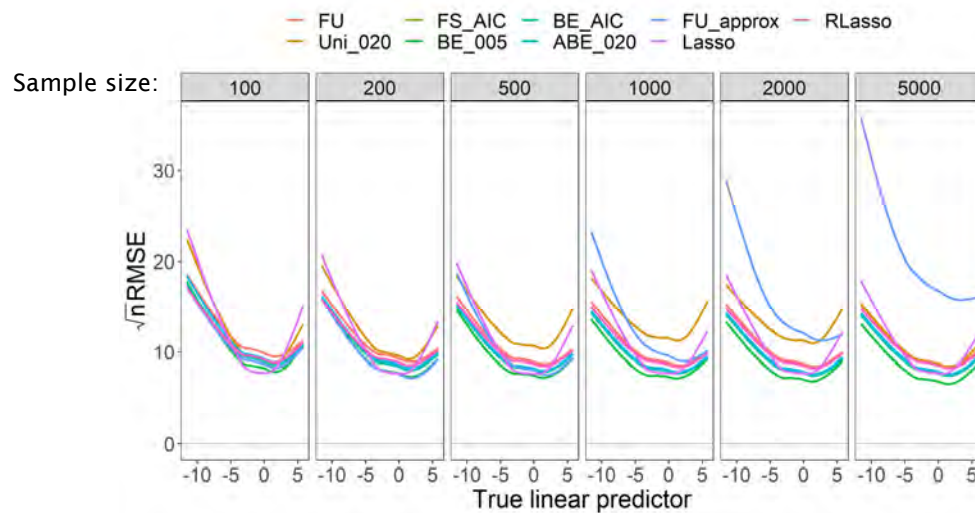




## Local bias of predictions



## Local RMSE of predictions



## Summary I

- Careful interpretation of conditional and unconditional performance!
- E.g. conditional coverage - not meaningful for variables selected in 5%.
- Variable selection methods have been described with 'bias away from zero', but this concerns the conditional bias only.
- Unconditionally, there is bias towards 0.
- Univariate filtering results strongly depend on correlation structure!

## Summary II


- For large samples ( $> 50$  NPV), BE(0.05) dominates all other methods in predictive accuracy.
- It is close to BIC – discover the true model if it is in the scope of models evaluated.
- BE works if true positive rate (TPR) is high for ‘true effects’ and false positive rate (FPR) is low for ‘null effects’.
- Therefore, variable inclusion frequencies (VIFs) may provide a guide towards whether we can trust the best BW model:
  - VIFs should be routinely computed and reported,
  - report also performance of ‘second-line’ models,
  - don’t trust a single model if selection is not sure.

## Summary III

- Forward selection inferior to backward elimination.
- Lasso performs well in the ‘center’, but shrinks towards the mean (pessimistic).
  - Problem probably estimation of penalty factor
- Lasso – problem with interpretability. (Remedy: ‘relaxed Lasso’)
- Background knowledge improves conditional measures and predictive accuracy because selection and estimation are disentangled.


## Summary IV


- Data-driven selection is a bad idea with small samples.
- Better to work with simple, defensible, fixed models.
- But: depends on modeling aim
  - Easy with descriptive models (modeler can ‚choose‘ the variables)
  - More difficult with prediction models
  - Explanatory models: ‚adjustment set‘ to minimize bias
    - but trade-off with MSE should be considered (near-instruments!)



A graphic illustration on a teal background. On the left, there are several white papers with lines of text, a magnifying glass with a black frame and an orange handle, a blue pencil, and a red sticky note with wavy lines. On the right, there is a yellow pencil. The text 'Case Study' is written in large, white, sans-serif font in the center-right.

from <https://www.packback.co/case-studies/>

 MEDICAL UNIVERSITY OF VIENNA

 STRATOS INITIATIVE

Georg Heinze, Christine Wallisch, Daniela Dunkler  
CeMSIS - Section for Clinical Biometrics

Part II-2

## Consulting situations



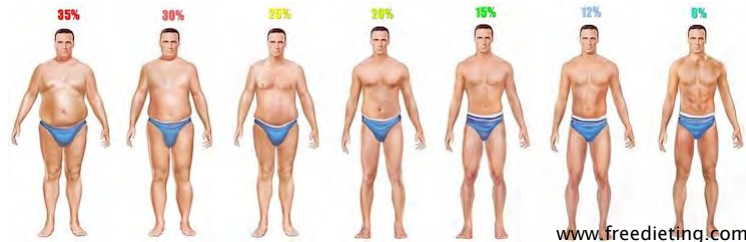
- 'We would like to approximate the proportion of body fat by simple anthropometric measures.'
- 'We want a prediction model for recurrent venous thromboembolism. Many risk factors were previously described, but the model should be clinically applicable for making therapy decisions. Can you please develop a parsimonious model?'
- 'We want a prediction model for survival after cervical cancer diagnosis. We know our predictors. There are only few events.'

## Case study 1: Body fat approximation

- Johnson's (1996) body fat data example
- Publicly available
- 252 men aged 21 to 81
- Response variable: % body fat (Siri formula), based on costly underwater density measurement
- Predictors: age, height, weight, 10 circumference measures
- First goal: approximation of % body fat



R markdown



www.freedieting.com

## Case study 2: Cervical cancer prognosis

- The request: 'We want a prediction model for survival after cervical cancer diagnosis. We know our predictors. There are only few events.'
- A retrospective cohort study:
  - 692 consecutive patients diagnosed with cervical cancer from two centers (Vienna, Innsbruck)
  - Median follow-up of 46 months



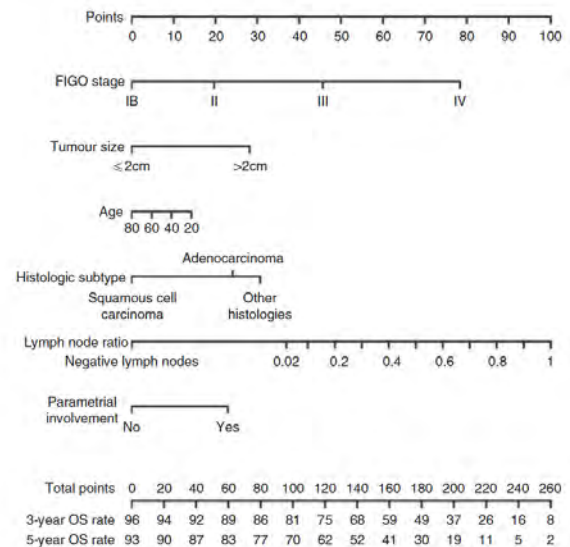
## Case study 2: cervical cancer prognosis

- Risk factors:
  - FIGO stage (I, II, III, IV) (3df)
  - Tumour size (<2cm, >2cm)
  - Age
  - Histologic subtype (squamous cell carcinoma, adenocarcinoma, other) (2df)
  - Proportion positive lymph nodes (2df)
  - Parametrical involvement (yes/no)
- These variables were available for 518 patients.
- 77 deaths → EPV=6.7

## Case study 2: cervical cancer prognosis

- critical EPV (6.7) → no variable selection
- L2-penalization (ridge regression)
- Presented as nomogram and web calculator.
- The nomogram shows the relative importance of the prognostic factors.

(Polterauer et al, 2012)



## Summary of case studies

- Variable selection should not always be considered (in particular, when EPV is very small).
- Stability investigations provide additional insights and should accompany variable selection.
- For explanatory models, use substance matter knowledge (or brains).

## Case study 1: Modeling body fat

### Aim of the analysis

We aim to develop a multivariable diagnostic prediction model for the approximation of percentage of body fat using simple anthropometric measurements and age. The primary objective is prediction and the secondary objective is the description of the adjusted association of each predictor with the outcome.

### Data dictionary and metadata

#### Introduction to the bodyfat data set

The original source of this data set is Roger W. Johnson (1996), "Fitting Percentage of Body Fat to Simple Body Measurements", Journal of Statistics Education, <http://jse.amstat.org/v4n1/datasets.johnson.html>. This data set contains the variables age, weight, height, ten body circumference measurements and estimates of the percentage of body fat determined by underwater weighing for 252 men.

#### Source data set

252 observations and 17 variables, no NAs

Name	Labels	Units	Measurement scale	Class	NAs
case	Case number			integer	0
brozek	Percent body fat using Brozek's equation	%	continuous	numeric	0
siri	Percent body fat using Siri's equation	%	continuous	numeric	0
density	Density determined from under water weighing	$gm/cm^3$	continuous	numeric	0
age	Age	years	continuous	numeric	0
weight	Weight	lbs	continuous	numeric	0
height	Height	inches	continuous	numeric	0
neck	Neck circumference	cm	continuous	numeric	0
chest	Chest circumference	cm	continuous	numeric	0
abdomen	Abdomen circumference	cm	continuous	numeric	0
hip	Hip circumference	cm	continuous	numeric	0
thigh	Thigh circumference	cm	continuous	numeric	0
knee	Knee circumference	cm	continuous	numeric	0
ankle	Ankle circumference	cm	continuous	numeric	0
biceps	Biceps (extended) circumference	cm	continuous	numeric	0
forearm	Forearm circumference	cm	continuous	numeric	0
wrist	Wrist circumference	cm	continuous	numeric	0

#### Data cleaning and working data set

An apparent error in height of case 42 was corrected. The implausible case 39 with weight > 300 kg was excluded. Units of weight and height were converted to kg and cm.

Hence, the working data set contains 251 observations and 17 variables, and no NAs.

Name	Labels	Units	Measurement scale	Class	NAs
case	Case number			integer	0
brozek	Percent body fat using Brozek's equation	%	continuous	numeric	0
siri	Percent body fat using Siri's equation	%	continuous	numeric	0
density	Density determined from under water weighing	$gm/cm^3$	continuous	numeric	0
age	Age	years	continuous	numeric	0
weight	Weight	kg	continuous	numeric	0
height	Height	cm	continuous	numeric	0
neck	Neck circumference	cm	continuous	numeric	0
chest	Chest circumference	cm	continuous	numeric	0
abdomen	Abdomen circumference	cm	continuous	numeric	0
hip	Hip circumference	cm	continuous	numeric	0
thigh	Thigh circumference	cm	continuous	numeric	0
knee	Knee circumference	cm	continuous	numeric	0
ankle	Ankle circumference	cm	continuous	numeric	0
biceps	Biceps (extended) circumference	cm	continuous	numeric	0
forearm	Forearm circumference	cm	continuous	numeric	0
wrist	Wrist circumference	cm	continuous	numeric	0

## Statistical analysis plan

### Statistical methods for main research aim

Linear regression will be used to model percent of body fat approximated by Siri's equation (outcome variable: `siri`). The following independent variables were considered:

- Age (years)
- Height (in cm)
- Weight (in kg)
- Neck circumference (in cm)
- Chest circumference (in cm)
- Abdomen circumference (in cm)
- Hip circumference (in cm)
- Thigh circumference (in cm)
- Knee circumference (in cm)
- Ankle circumference (in cm)
- Biceps circumference (in cm)
- Forearm circumference (in cm)
- Wrist circumference (in cm)

Domain expertise:

We consider height and abdomen circumference as pivotal in the estimation of bodyfat. Hence these two variables should not be subjected to variable selection but rather always be included in the models.

We will fit several models:

- We will start with a model containing all candidate predictors (the global basic model) collected in the original data set,
- Variable selection, in particular, backward elimination with AIC as stopping criterion will be applied to reduce the number of predictors in order to obtain a parsimonious model for application (BE selected basic model),

- As sensitivity analysis we will conduct augmented backward elimination (ABE) with default values of the hyperparameters (ABE selected basic model).
- As alternative approach, we will fit a model based on the idea of dimensionality reduction (DR) as outlined in Burnham & Anderson (2002) to address the expected multicollinearity of the anthropometric measurements (global, BE and ABE selected DR model).

For selected models, stability will be evaluated by computing model selection frequencies and variable inclusion frequencies using subsampling with a fraction of 0.5, and root mean squared difference ratio (RMSDR) and relative conditional bias using the nonparametric bootstrap. Sampling variability of regression coefficients will be assessed by the 2.5th and 97.5th percentiles of the bootstrapped coefficients, considering coefficients of unselected predictors as 0.

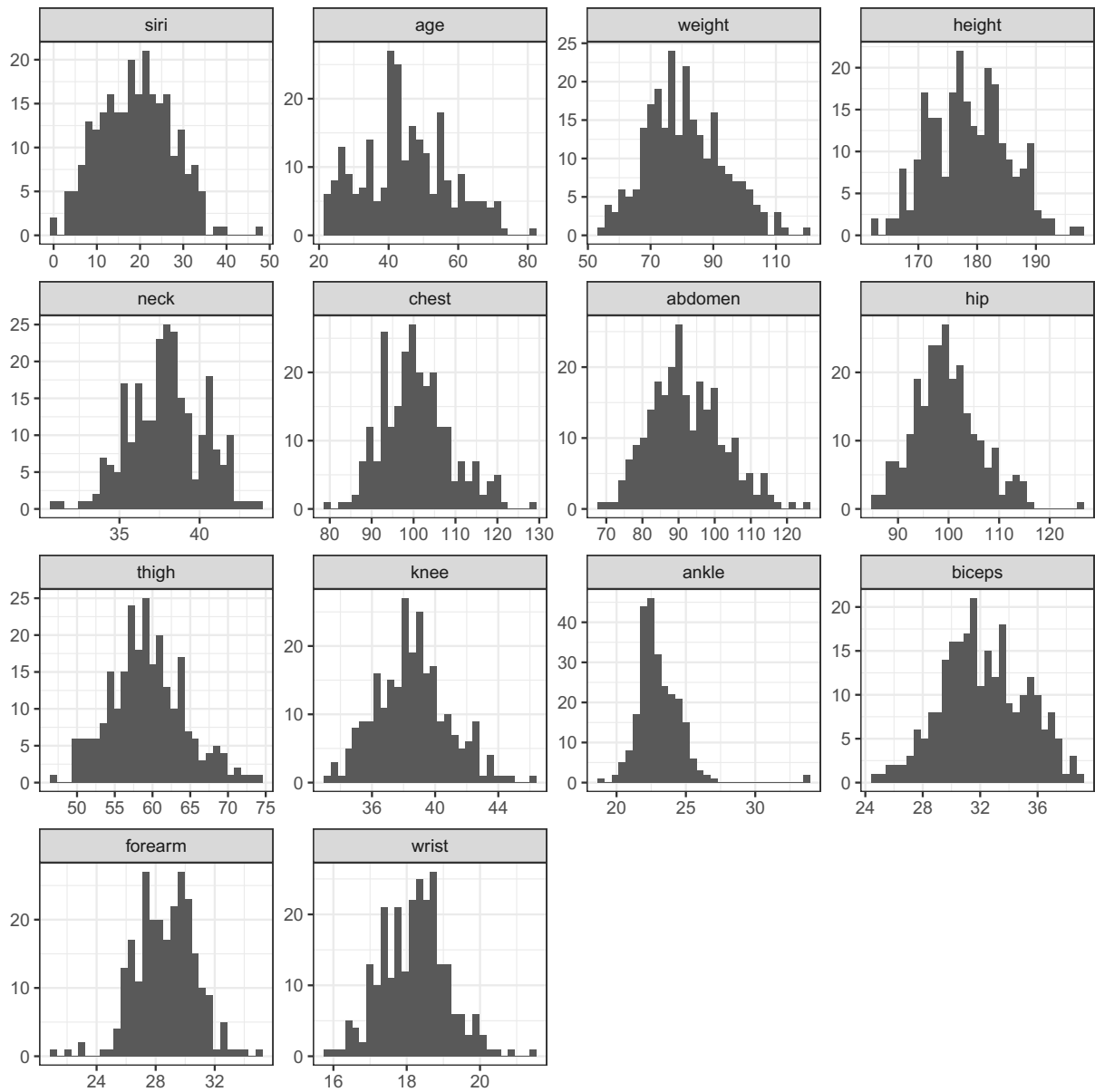
In a data screening step preceding modeling, we will investigate univariate distributions of all predictors and the outcome variable, and the correlation between the predictor variables.

## Data screening

### Univariate distributions of predictors and the outcome variable

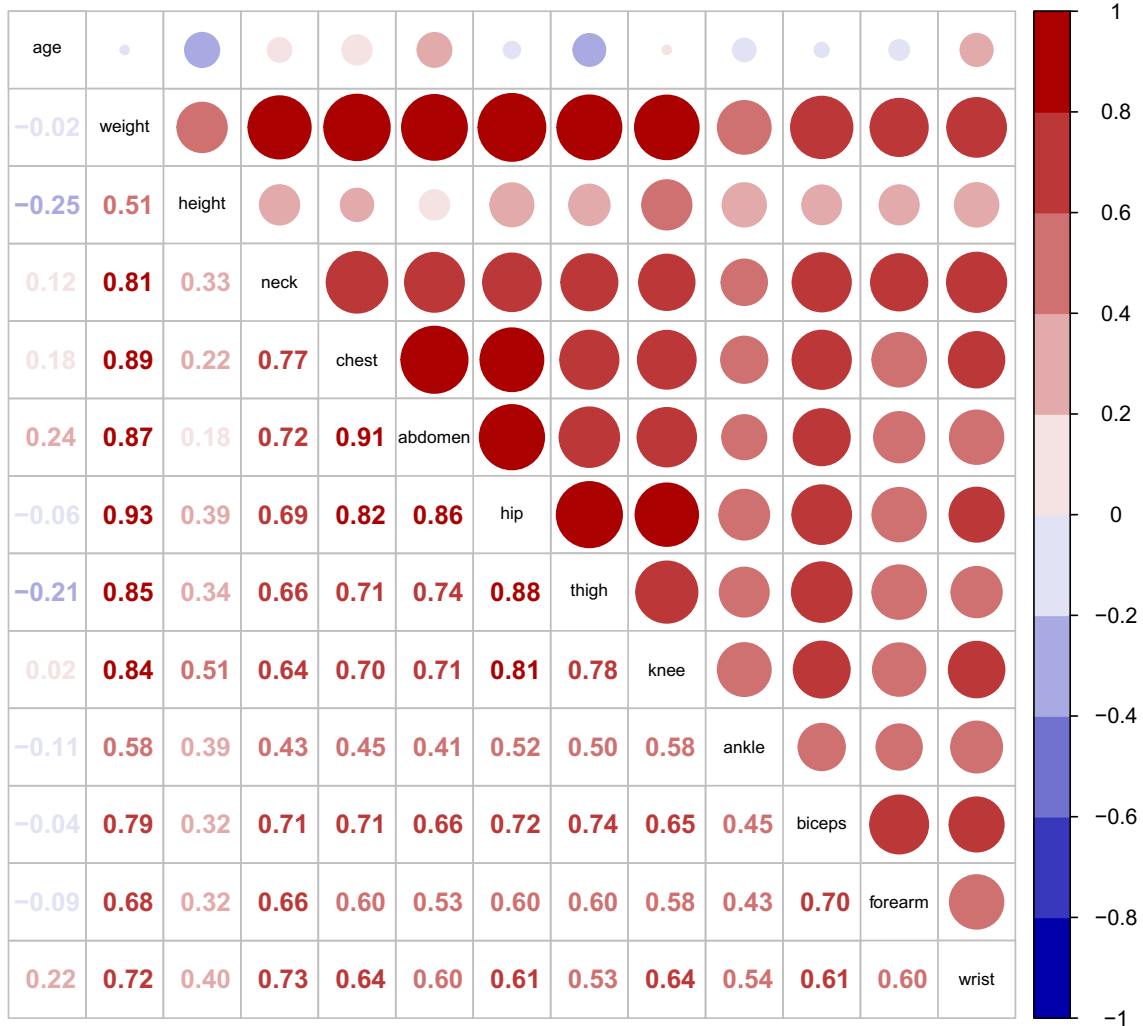
Table 1: Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 50	Pctl. 75	Max
siri	251	19.1	8.3	0	12.4	19.2	25.2	47.5
age	251	44.9	12.6	22	35.5	43	54	81
weight	251	80.9	12.3	53.8	72.1	80	89.4	119.3
height	251	178.6	6.6	162.6	173.4	177.8	183.5	197.5
neck	251	37.9	2.3	31.1	36.4	38	39.4	43.9
chest	251	100.7	8.1	79.3	94.3	99.6	105.3	128.3
abdomen	251	92.3	10.2	69.4	84.5	90.9	99.2	126.2
hip	251	99.7	6.5	85	95.5	99.3	103.3	125.6
thigh	251	59.3	5	47.2	56	59	62.3	74.4
knee	251	38.5	2.3	33	37	38.5	39.9	46
ankle	251	23.1	1.6	19.1	22	22.8	24	33.9
biceps	251	32.2	2.9	24.8	30.2	32	34.3	39.1
forearm	251	28.7	2	21	27.3	28.7	30	34.9
wrist	251	18.2	0.9	15.8	17.6	18.3	18.8	21.4



The distribution of most variables are approximately symmetric. Only some measurements of ankle and hip are very high, but are still considered plausible.

### Bivariate Pearson correlation analysis



Out of the 13 predictor variables, there are two pairs of variables exhibiting Pearson correlation coefficients greater than 0.9 (hip with weight, abdomen with chest), and a group of ten variables with all pairwise correlation coefficients greater than 0.5 (forearm, biceps, wrist, neck, knee, hip, weight, thigh, abdomen, chest). These high correlations impose some challenges in model development and interpretation.

Interpretation of non-selected variables as ‘nonpredictive’ is highly problematic (both under the full model and submodel views).



## Model building

Some statisticians recommended 15 observations per (design) variable as the minimum to obtain a statistical model with adequate accuracy.

The number of observations per variable in our setting is  $251/13=19$ .

Here, the number of variables actually corresponds to the number of design variables in the model. Thus, if we included further terms to address non-linearities (e.g. by use of splines or fractional polynomials) or interactions, the number of observations per variable would decrease correspondingly.

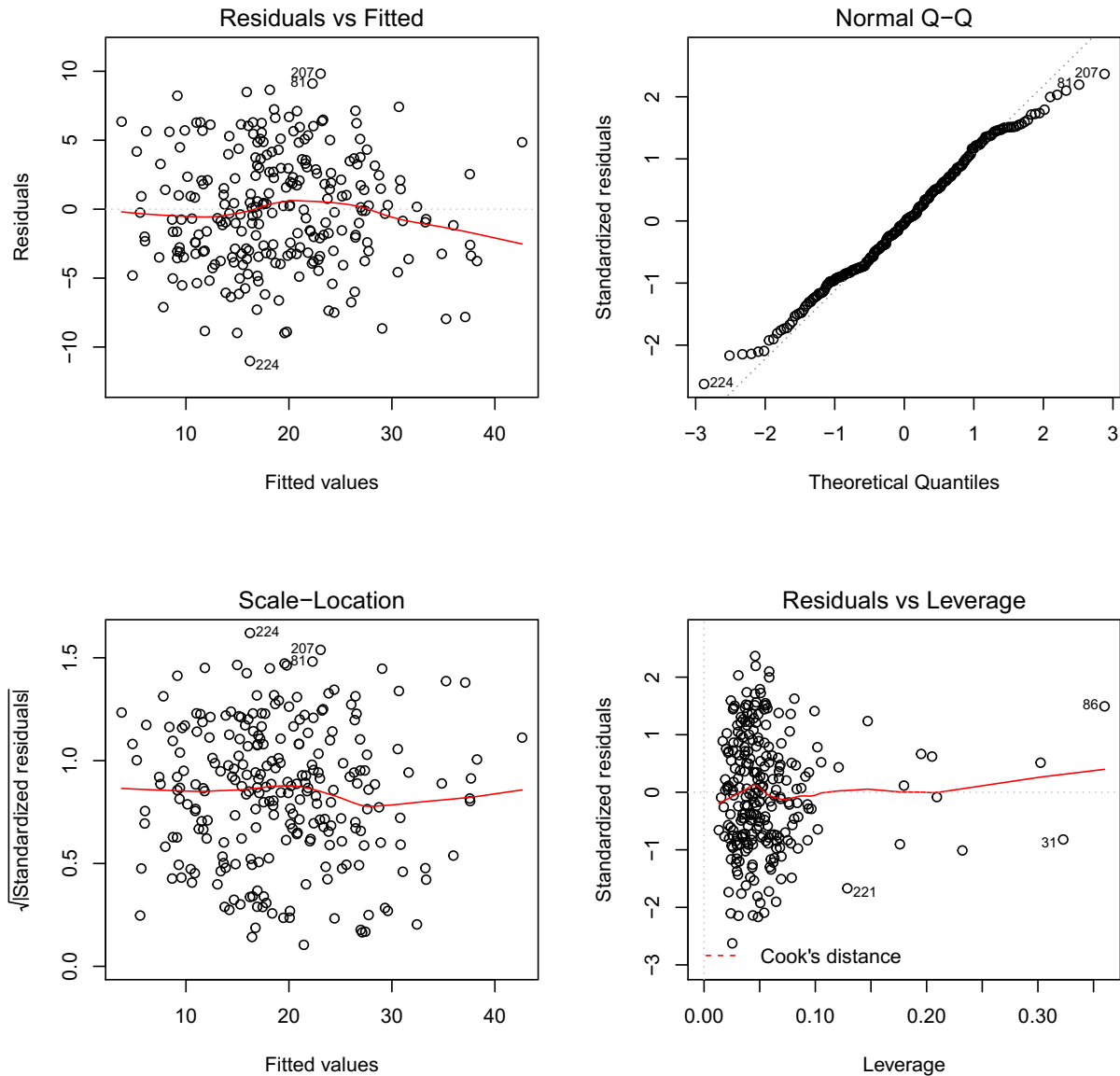
## Global basic model

As first step, we estimate the global model including all predictors.

```
##
## Call:
## lm(formula = formula, data = bodyfat, x = T, y = T)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.0149  -3.1706  -0.1178   3.0133   9.8257
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.34244    23.32885   0.058 0.954160
## age          0.07380     0.03191   2.313 0.021595 *
## weight      -0.04110     0.14758  -0.279 0.780863
## height      -0.09800     0.07512  -1.305 0.193302
## neck        -0.39426     0.23406  -1.684 0.093414 .
## chest       -0.11906     0.10808  -1.102 0.271764
## abdomen     0.90082     0.09098   9.901 < 2e-16 ***
## hip         -0.14603     0.14356  -1.017 0.310112
## thigh       0.17805     0.14629   1.217 0.224754
## knee        -0.04099     0.24505  -0.167 0.867287
## ankle       0.18549     0.21951   0.845 0.398952
## biceps      0.17760     0.17008   1.044 0.297457
## forearm     0.27722     0.20659   1.342 0.180914
## wrist       -1.83017     0.52940  -3.457 0.000647 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.249 on 237 degrees of freedom
## Multiple R-squared:  0.753, Adjusted R-squared:  0.7394
## F-statistic: 55.57 on 13 and 237 DF, p-value: < 2.2e-16
```

## Model diagnostics

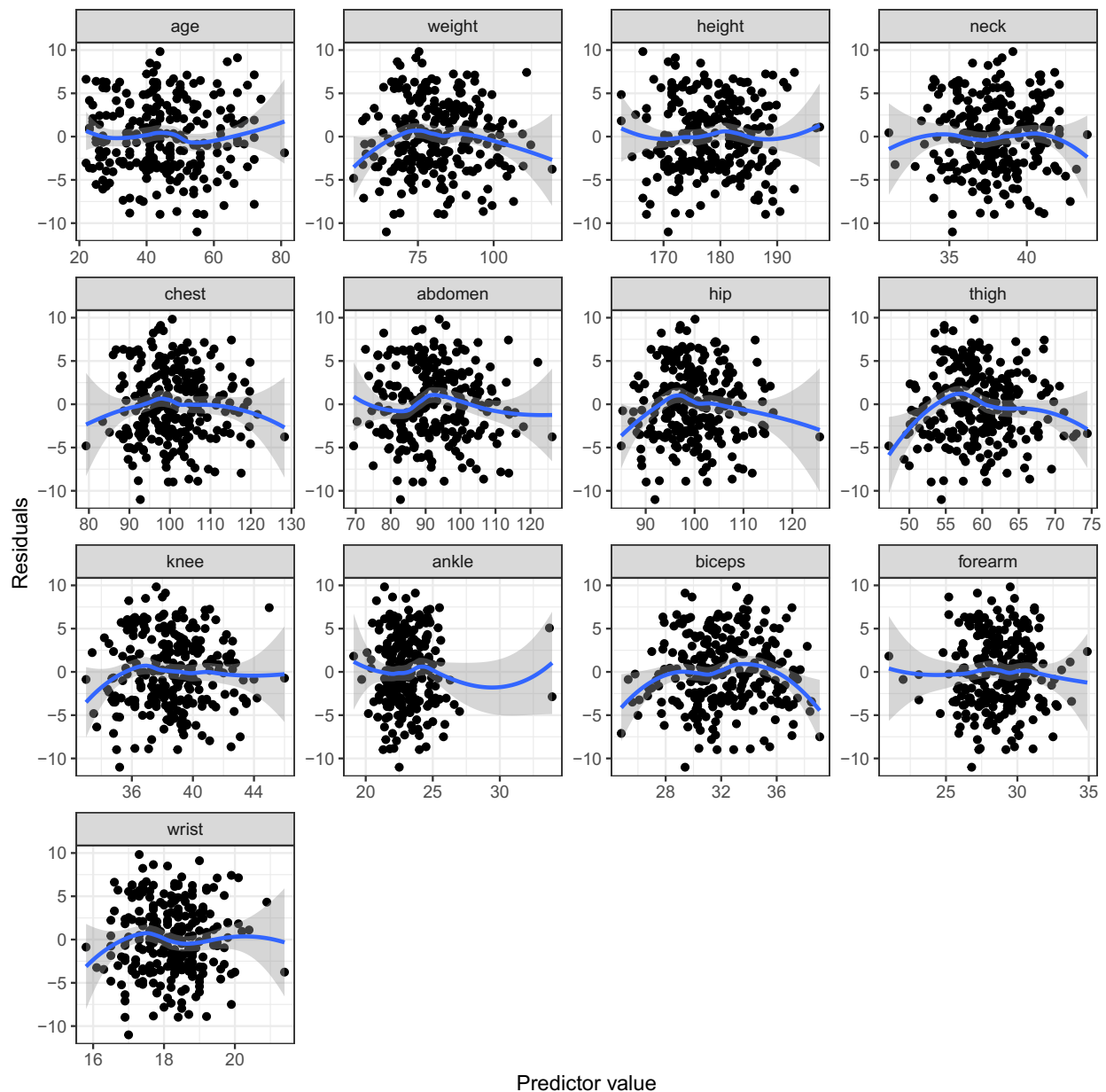
We call the standard residual diagnostics available for `lm` objects:



These plots can be used to assess internal validity of the model. The first plot does not give rise to concerns on local biases, i.e., the expected values of the residuals are 0 over the full range of predictions. The second plot confirms approximate normal distribution of residuals, which is a prerequisite for the interpretability of confidence intervals and p-values, but actually not an assumption of the model. The scale-location plot suggests that the residuals have a constant variance (absence of heteroscedasticity).

The fourth plot suggests that some points may have disproportional impact on the regression results. However, by checking the plausibility of the predictors' distributions in the data screening step, we would not decide to omit such observations. 'Robust statistics' were developed to downweight such influential points in the analysis in order to reduce their influence. As we will demonstrate later, we will choose another route for robustifying our results.

Residuals plotted against predictor values inform about possibly violated linearity assumptions of effects:



Some variables, e.g. hip, thigh and biceps, show some evidence for misspecification of the functional form of association with the outcome. In this course, we will not deal with functional form selection. Our approach to deal with the problem is to robustify the model by computing a robust sandwich variance. In this way, our model is still easily interpretable as summary analysis informing about the **average linear effect of each variable** adjusted for the others, with the misspecification adequately reflected by possibly enlarged robust standard errors and confidence intervals.

In fact, in this example the robust standard errors are partly even smaller than their model-based counterparts:

```
##           Coef      SE robust SE SE ratio
## (Intercept)  1.34 23.33      23.09      0.99
## age          0.07  0.03       0.03      0.87
## weight      -0.04  0.15       0.14      0.98
## height      -0.10  0.08       0.07      0.97
## neck        -0.39  0.23       0.21      0.92
## chest       -0.12  0.11       0.10      0.97
## abdomen     0.90  0.09       0.08      0.92
## hip         -0.15  0.14       0.13      0.91
## thigh       0.18  0.15       0.13      0.88
## knee        -0.04  0.25       0.22      0.90
## ankle       0.19  0.22       0.21      0.98
## biceps      0.18  0.17       0.16      0.91
## forearm     0.28  0.21       0.14      0.70
## wrist      -1.83  0.53       0.47      0.89
```

Since the predictors are highly correlated, one could argue that not all predictors are needed for accurate prediction of bodyfat percentage. This justifies to apply variable selection to reduce the set of predictors.

## BE selected basic model

We did not subject height and abdomen to variable selection since we believe that they play a central role for estimating the proportion of body fat. All other variables are considered competitively for estimation. Backward elimination with AIC as stopping criterion was chosen as variable selection algorithm since AIC is an appropriate criterion for fitting prediction models.

```
##
## Call:
## lm(formula = siri ~ age + height + neck + chest + abdomen + forearm +
##      wrist, data = bodyfat, x = T, y = T)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.5596  -3.1079  -0.1909   3.1729   9.4388
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.45437     8.17898   0.667 0.505484
## age          0.06086     0.02480   2.454 0.014820 *
## height      -0.12631     0.04749  -2.659 0.008348 **
## neck        -0.33160     0.21892  -1.515 0.131144
## chest       -0.13339     0.08762  -1.522 0.129230
## abdomen     0.87380     0.06483  13.478 < 2e-16 ***
## forearm     0.36215     0.19191   1.887 0.060335 .
## wrist      -1.73444     0.48427  -3.582 0.000412 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.236 on 243 degrees of freedom
## Multiple R-squared:  0.7482, Adjusted R-squared:  0.741
## F-statistic: 103.2 on 7 and 243 DF,  p-value: < 2.2e-16
```

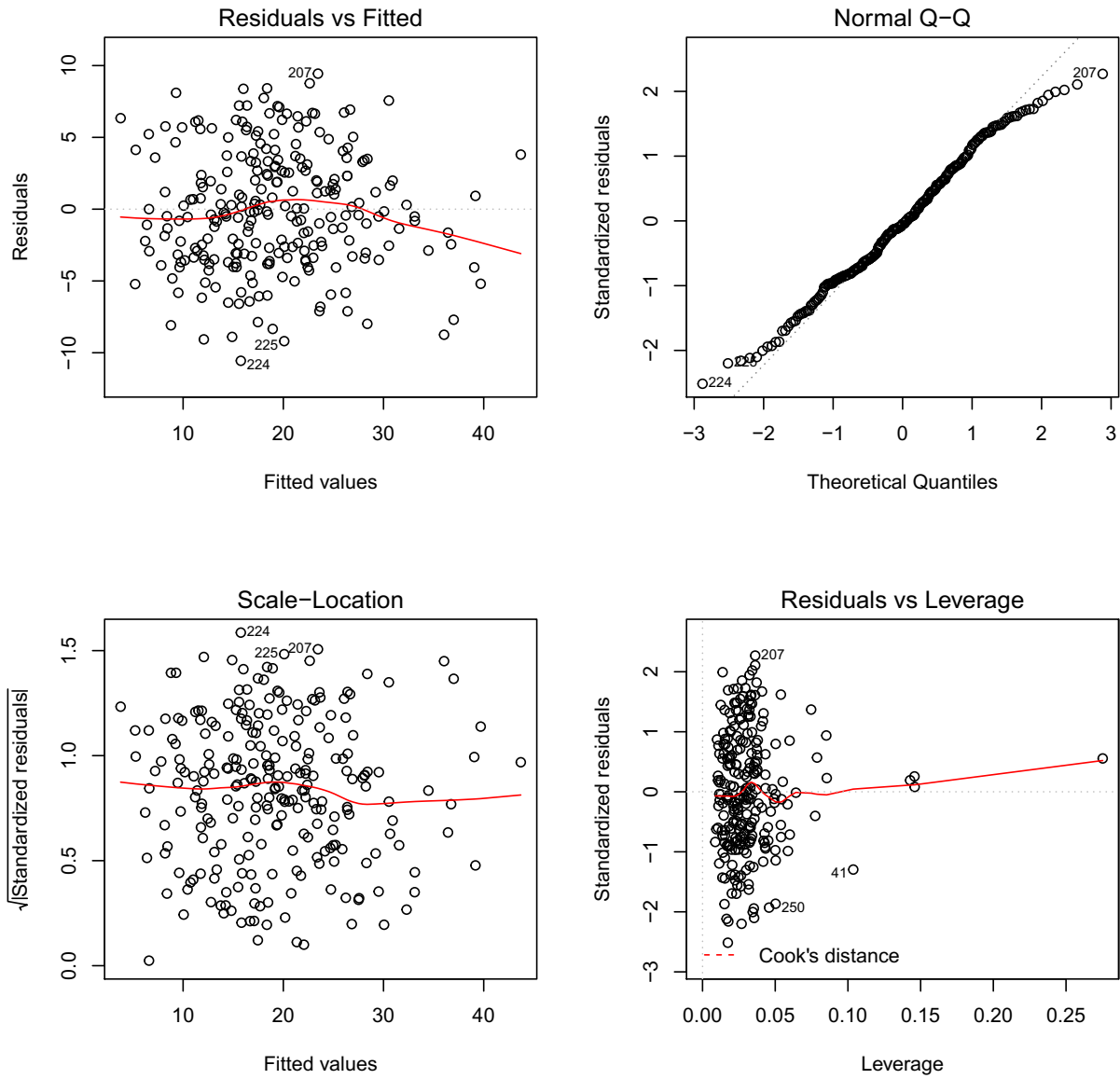
The adjusted  $R^2$  in the global model increased only slightly from 0.739 to 0.741 in the selected model.

Again, most robust standard errors are slightly lower:

##	Coef	SE	robust SE	SE ratio
## (Intercept)	5.45	8.18	8.21	1.00
## age	0.06	0.02	0.02	0.91
## height	-0.13	0.05	0.05	1.05
## neck	-0.33	0.22	0.20	0.91
## chest	-0.13	0.09	0.09	0.97
## abdomen	0.87	0.06	0.06	0.96
## forearm	0.36	0.19	0.14	0.73
## wrist	-1.73	0.48	0.44	0.92

### Model diagnostics

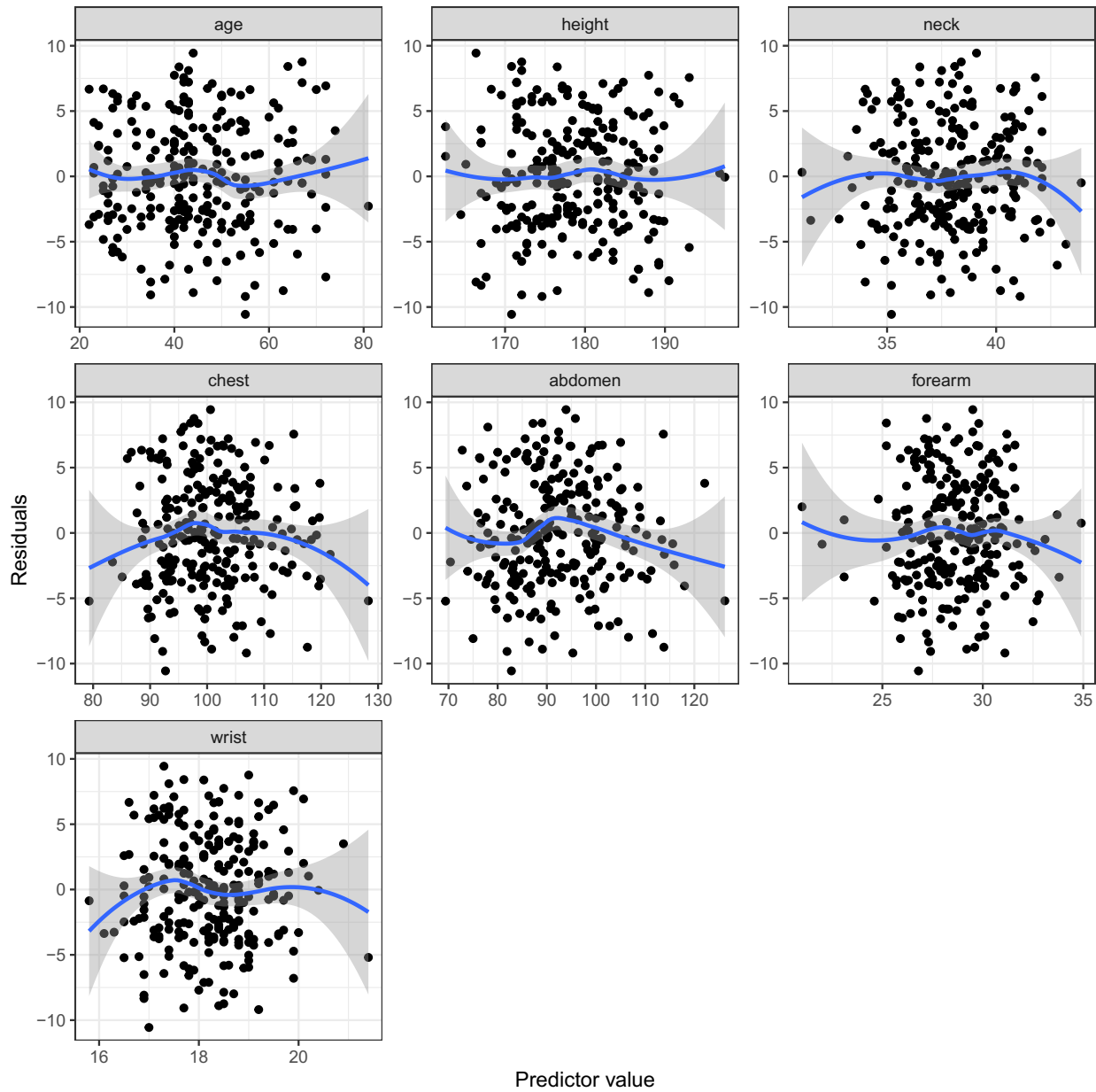
Again, we perform model diagnostics by means of some plots of residuals.



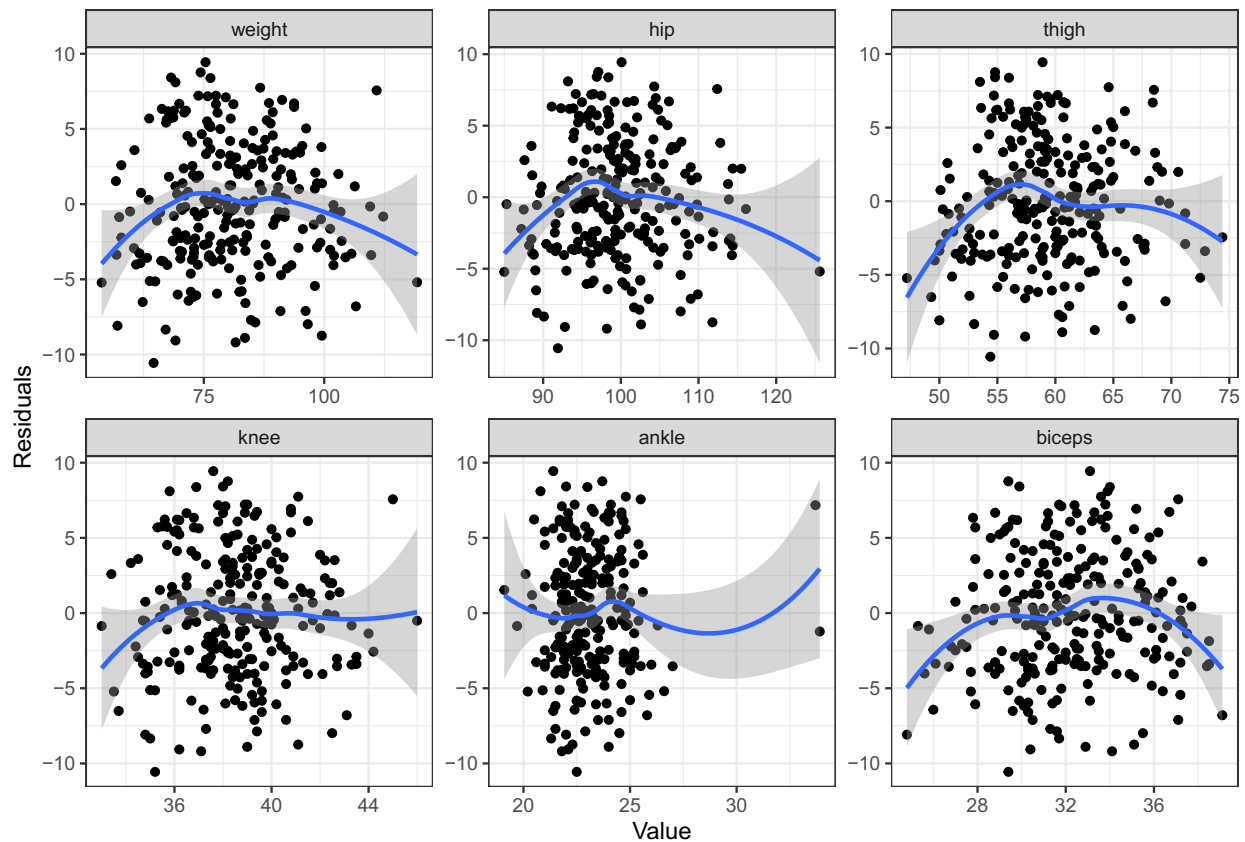
These plots lead to similar conclusions as for the global model.

In the following, we plot residuals with loess smoothers against the values of the predictor variables and against the variables which were eliminated from the model.

Selected variables



## Non-selected variables



The residuals are hardly systematically related with any of the selected predictors. However, there is some inversely U-shaped association with weight, hip, thigh and biceps. These variables may need consideration with a quadratic or even more complex nonlinear functional form in order to improve the model, and were probably eliminated because their functional form was misspecified. As already discussed, we do not treat this in this course in more detail. We rather apply robust standard errors to account for model misspecification.



## Stability of the BE selected basic model

By default, the output of software performing variable selection does not inform about the instability of models and the additional uncertainty that is incurred by selection. Hence, we calculate stability measures to investigate this. We demonstrate how to compute and interpret variable inclusion frequencies (VIF), model selection frequencies (MSF), pairwise inclusion frequencies, the relative conditional bias (RCB) and the root mean squared difference ratio (RMSD ratio).

Estimands of VIF, MSF, RCB and RMSD ratio are explained in our paper ‘Selection of variables for multivariable models: Opportunities and limitations in quantifying model stability by resampling’ (Wallisch et al, *Statistics in Medicine* 2021). In short,

- The variable inclusion frequency quantifies how likely a variable is selected with a random sample of given size from the population and applying a specific variable selection algorithm.
- The model selection frequency indicates how likely a specific combination of variables are selected.
- Pairwise inclusion frequencies quantify how likely pairs of variables are selected. They inform about ‘rope teams’ and ‘competitors’ among variables.
- The relative conditional bias expresses the bias that is introduced into regression coefficients by applying variable selection relative to the (assumed unbiased) global model.
- Finally, the RMSD ratio quantifies inflation or deflation of standard errors caused by applying variable selection.

The stability measures VIF, MSF and the pairwise inclusion frequencies are calculated based on 500 samples of size  $N/2$  (with  $N$  denoting the sample size) drawn without replacement (subsampling) whereas RCB and RMSD ratio are calculated based on 500 bootstrap samples drawn with replacement according to the recommendations of Wallisch et al. (2021). (We restrict to 500 resamples to speed up computations. In real analyses, of course higher numbers of resamples can be used.)

### Summary of the model stability

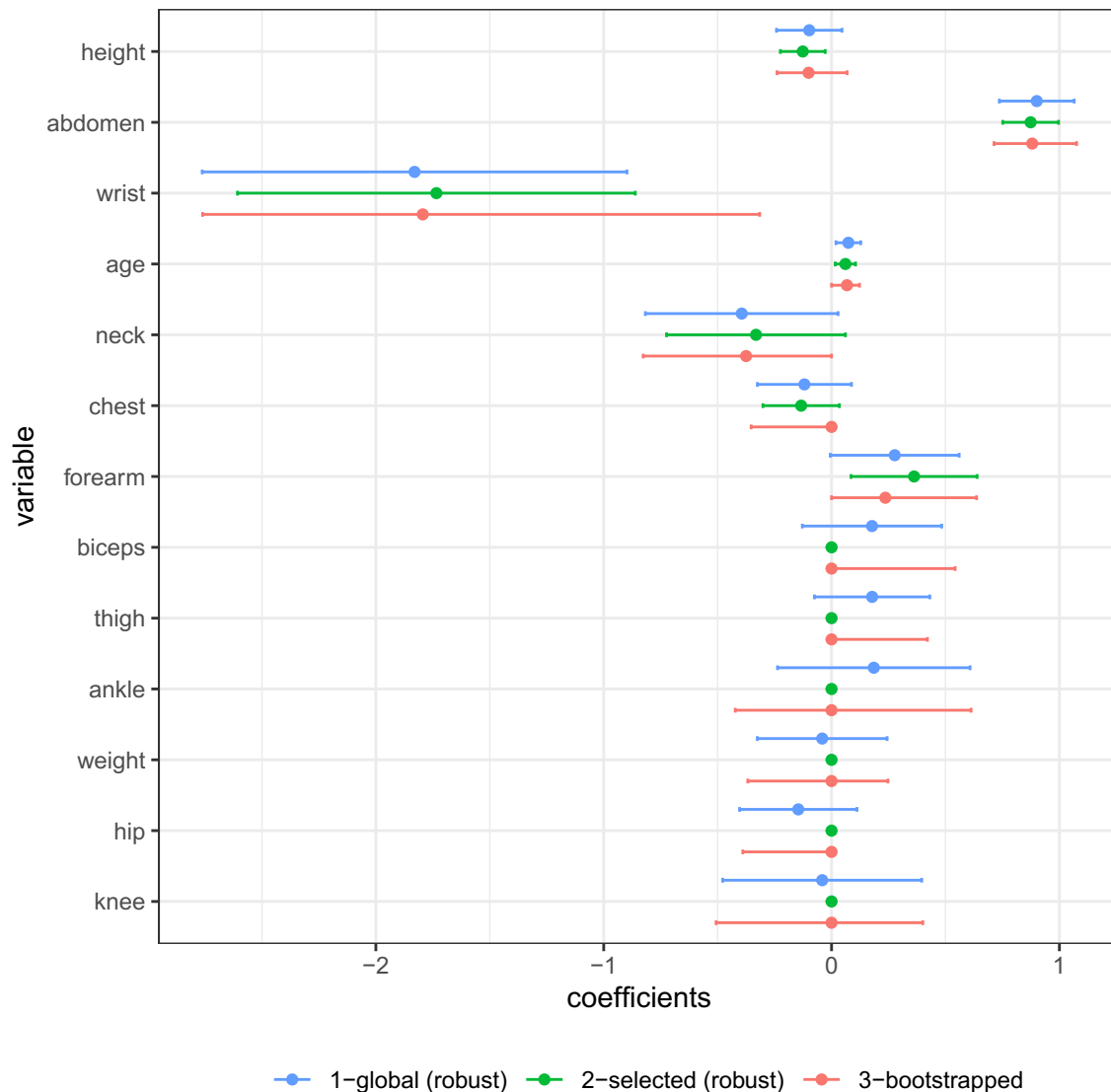
Below we report the coefficients from the global and the selected model, the bootstrapped sampling distributions of coefficients, the corresponding VIFs, RCBs and RMSD ratios.

##	Coef_global	SE_rob_global	Coef_sel	SE_rob_sel	VIF	RCB	RMSD	ratio
## height	-0.10	0.07	-0.13	0.05	1.00	1.18		1.05
## abdomen	0.90	0.08	0.87	0.06	1.00	-1.60		1.02
## wrist	-1.83	0.47	-1.73	0.44	0.91	-1.84		1.05
## age	0.07	0.03	0.06	0.02	0.65	3.53		1.17
## neck	-0.39	0.21	-0.33	0.20	0.37	29.84		1.23
## chest	-0.12	0.10	-0.13	0.09	0.32	73.49		1.09
## forearm	0.28	0.14	0.36	0.14	0.29	44.72		1.15
## biceps	0.18	0.16	NA	NA	0.27	97.48		1.12
## thigh	0.18	0.13	NA	NA	0.25	50.43		1.07
## ankle	0.19	0.21	NA	NA	0.21	84.65		1.16
## weight	-0.04	0.14	NA	NA	0.20	281.75		0.98
## hip	-0.15	0.13	NA	NA	0.19	78.72		1.04
## knee	-0.04	0.22	NA	NA	0.08	261.03		0.69

- VIFs: As predefined, height and abdomen have a selection frequency of 100% in the resampling models. Also wrist was selected in 91% of the models. However, many predictors were selected less frequently due to high collinearity. Surprisingly, weight seems rather unimportant with a VIF around 20%.

- RCB: this measure is given in per cent. Very small biases are exhibited by the top selected variables, up to age. For these variables we can safely ignore overestimation effects due to selection. Variables for which selection is less sure show severe overestimation if selected.
- RMSD ratio: These measures suggest that only for weight and knee, the uncertainty in the estimation reduces by applying data-driven variable selection. For all other variables, the uncertainties induced by selection add up to the model uncertainty and this finally gives larger errors than if the global model was prespecified. However, our study (Wallisch 2021) revealed that with high correlation between the predictor variables, the RMSD ratios could be overestimated by 10-20%, such that the variance inflation is probably ignorable even for neck and age. In fact, our study revealed that the variance inflation is hard to estimate by resampling, and unbiased estimates can only be obtained with orthogonal predictors.

Here we illustrate the bootstrapped sampling distribution of the selected coefficients. Generally, the bootstrap 'confidence' intervals are wider than their robust counterparts, reflecting additional variability by considering the selection step as stochastic rather than conditioning on the selected model. Some bootstrap point estimates are exactly 0 as they are medians of the selected coefficients.



### Model selection frequency (MSF)

Here is the top-10 list of selected models during the resampling procedure:

```
##
##           (Intercept)+age+height+abdomen+wrist
##                                     0.072
## (Intercept)+age+height+chest+abdomen+forearm+wrist
##                                     0.030
##           (Intercept)+age+height+abdomen+ankle+wrist
##                                     0.028
## (Intercept)+age+height+chest+abdomen+biceps+wrist
##                                     0.028
## (Intercept)+age+height+neck+abdomen+forearm+wrist
##                                     0.028
## (Intercept)+age+height+chest+abdomen+ankle+biceps+wrist
##                                     0.024
##           (Intercept)+height+abdomen+wrist
##                                     0.024
## (Intercept)+age+height+chest+abdomen+wrist
##                                     0.022
## (Intercept)+age+height+neck+abdomen+biceps+wrist
##                                     0.022
## (Intercept)+age+height+neck+abdomen+thigh+wrist
##                                     0.022
```

The MSFs are very low, and our selected model

[1] “(Intercept)+age+height+neck+chest+abdomen+forearm+wrist”  
with a MSF of 0.008 can only be found at position 30 on the list.

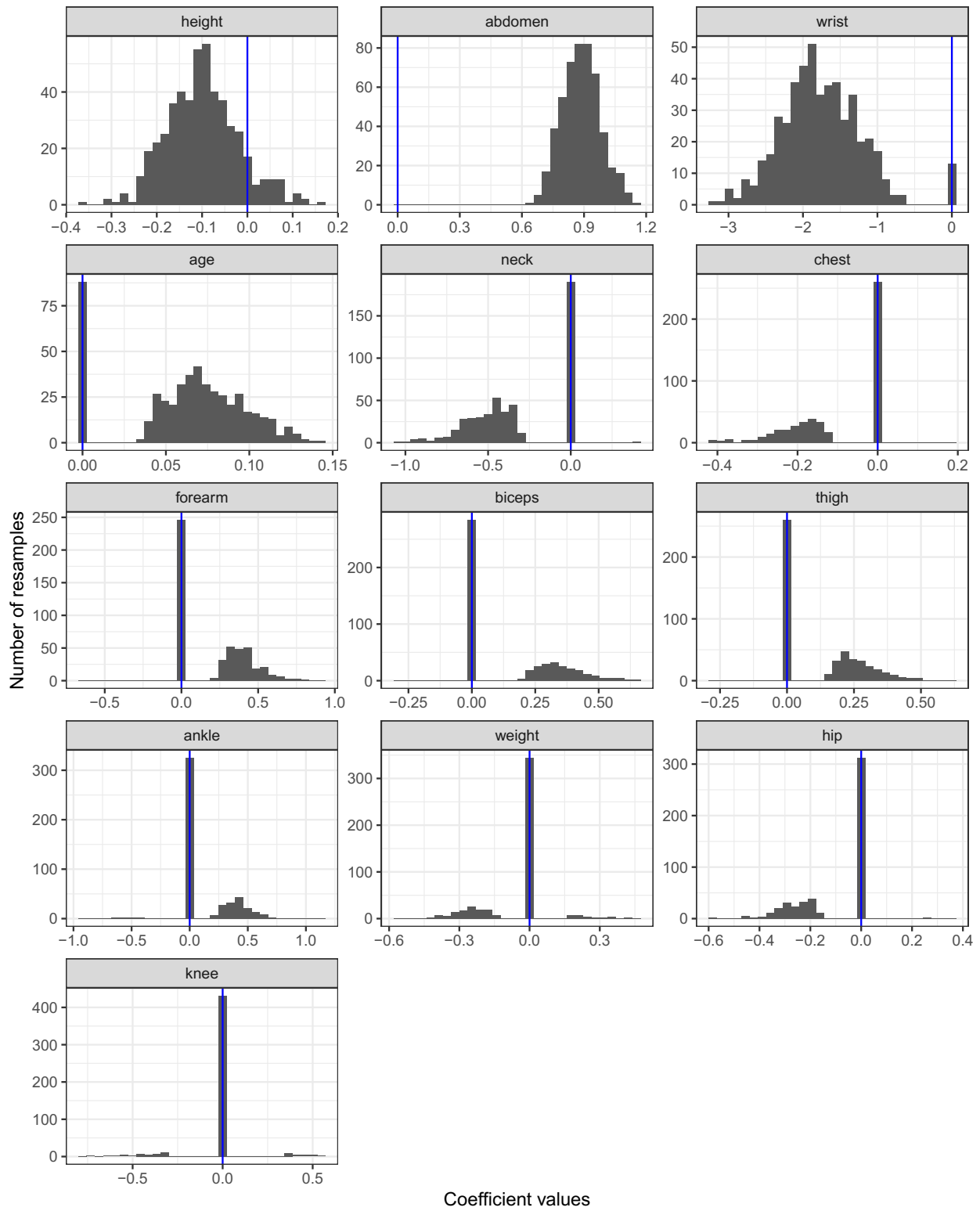
### Pairwise inclusion frequencies (PIF)

Pairwise inclusion frequencies inform about “rope teams” and “competitors” among the predictors. The following table shows pairwise PIFs and VIFs on the diagonal. For example, thigh and biceps were both selected in only 3.2% of the resamples, while one would expect a frequency of 6.9% ( $= 27.4\% \times 25.2\%$ ) given independent selection. In this table, we used significance of a  $\chi^2$  test at the 0.01 level as the formal criterion for the flags in the lower triangle. Therefore, the pair (thigh, biceps) is flagged with “-” in the lower triangle of the matrix below. Thigh and hip are flagged with “+” because they are simultaneously selected in 6.2% of the resamples, while the expectation under independence is only 4.8%. Interestingly, age forms a “rope team” with forearm, but is a competitor to thigh, ankle and weight. Variables with VIF of 100% were not considered in this table.

##		wrist	age	neck	chest	forearm	biceps	thigh	ankle	weight	hip	knee
##	wrist	91.4	64.2	31	30.4	27.2	26.6	23.6	20	18.2	17.2	7.2
##	age	+	64.8	24	21.6	21.4	19.4	20.6	15.8	9	10.6	4.4
##	neck	+		37.2	7.6	14	11.2	11.4	4.4	3.4	10.6	2.8
##	chest			-	32.4	10.2	12.2	5	7.8	1.6	6.2	2.6
##	forearm		+	+		28.8	4.2	9.8	4.4	6.8	6.6	2.2
##	biceps	+			+	+	27.4	3.2	6	6	6.2	3
##	thigh		-		-	+	-	25.2	5.2	8.2	10.2	2.6
##	ankle		-	-					20.6	5.6	3.6	1.6
##	weight		-	-	-					20.2	2.6	2.6
##	hip			-				+			19.2	1
##	knee											7.8

### Resampling distribution of predictors

All variables except for height and abdomen, which were forced into the model, have a spike at zero in the resampling distribution of their coefficients. If age was selected, it clearly had a positive effect on bodyfat. For some other variables, e.g., chest, weight, thigh, weight, hip and knee, both negative and positive coefficients were observed in the resampled models. The selection of those coefficients, and also the sign their coefficients obtained, strongly depended on the selected companion predictors.



### Sensitivity analysis: fitting the ABE selected basic model

As sensitivity check, we fitted a model with augmented backward elimination (ABE) and  $\tau = 0.05$ , which only excludes a variable if all other coefficients only change by less than 5% (ABE selected basic model).

Applying ABE to the set of predictors leads to a very large model. Only the variable knee was excluded whereas backward elimination excluded several variables. This finding points to instability of the BE selected basic model and probably a greater stability of the ABE selected basic model.

```
##
## t test of coefficients:
##
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.084268  22.915848  0.0473 0.9623017
## age          0.072416  0.025959  2.7896 0.0057040 **
## weight      -0.044278  0.141261 -0.3134 0.7542148
## height      -0.100232  0.072900 -1.3749 0.1704489
## neck        -0.390332  0.211217 -1.8480 0.0658408 .
## chest       -0.118194  0.104597 -1.1300 0.2596143
## abdomen     0.901160  0.083130 10.8403 < 2.2e-16 ***
## hip         -0.147901  0.130752 -1.1312 0.2591255
## thigh       0.170982  0.127670  1.3393 0.1817668
## ankle       0.178756  0.211569  0.8449 0.3990124
## biceps      0.179022  0.155346  1.1524 0.2503105
## forearm     0.275360  0.141427  1.9470 0.0527093 .
## wrist      -1.835837  0.473665 -3.8758 0.0001375 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## Dimensionality reduction (DR) approach: model building with combined variables

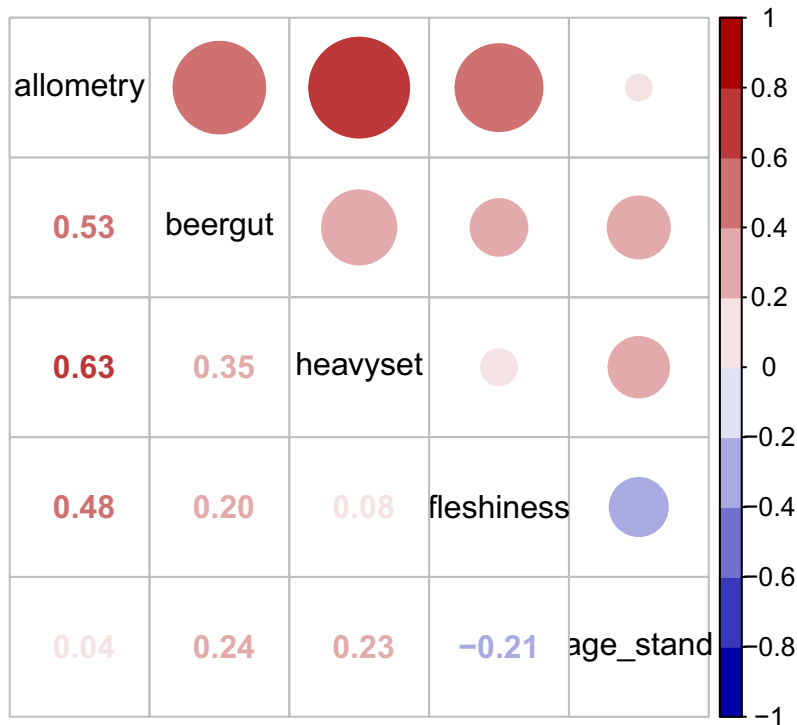
Because of the high correlation between the predictors, only small changes in the data may have tremendous effects on the estimated coefficients. One way to deal with this problem is to derive summary variables and use them for modeling the outcome.

Such an approach was pursued in the book of Burnham & Anderson (2002) for the estimation of bodyfat. For this model, the following new variables were computed (see the book for further explanation on how these summaries were justified):

- allometry =  $\log(\text{weight})/\log(\text{height})$
- beergut =  $\text{abdomen}/\text{chest}$
- heavysset =  $(\text{knee} * \text{wrist} * \text{ankle})^{1/3}/\text{height}$
- fleshiness =  $(\text{biceps} * \text{thigh} * \text{forearm}/(\text{knee} * \text{wrist} * \text{ankle}))^{1/3}$

In addition, age was standardized.

The new variables should express the major dimensions of anthropometry. This is confirmed by much lower correlation coefficients:



## Global DR model

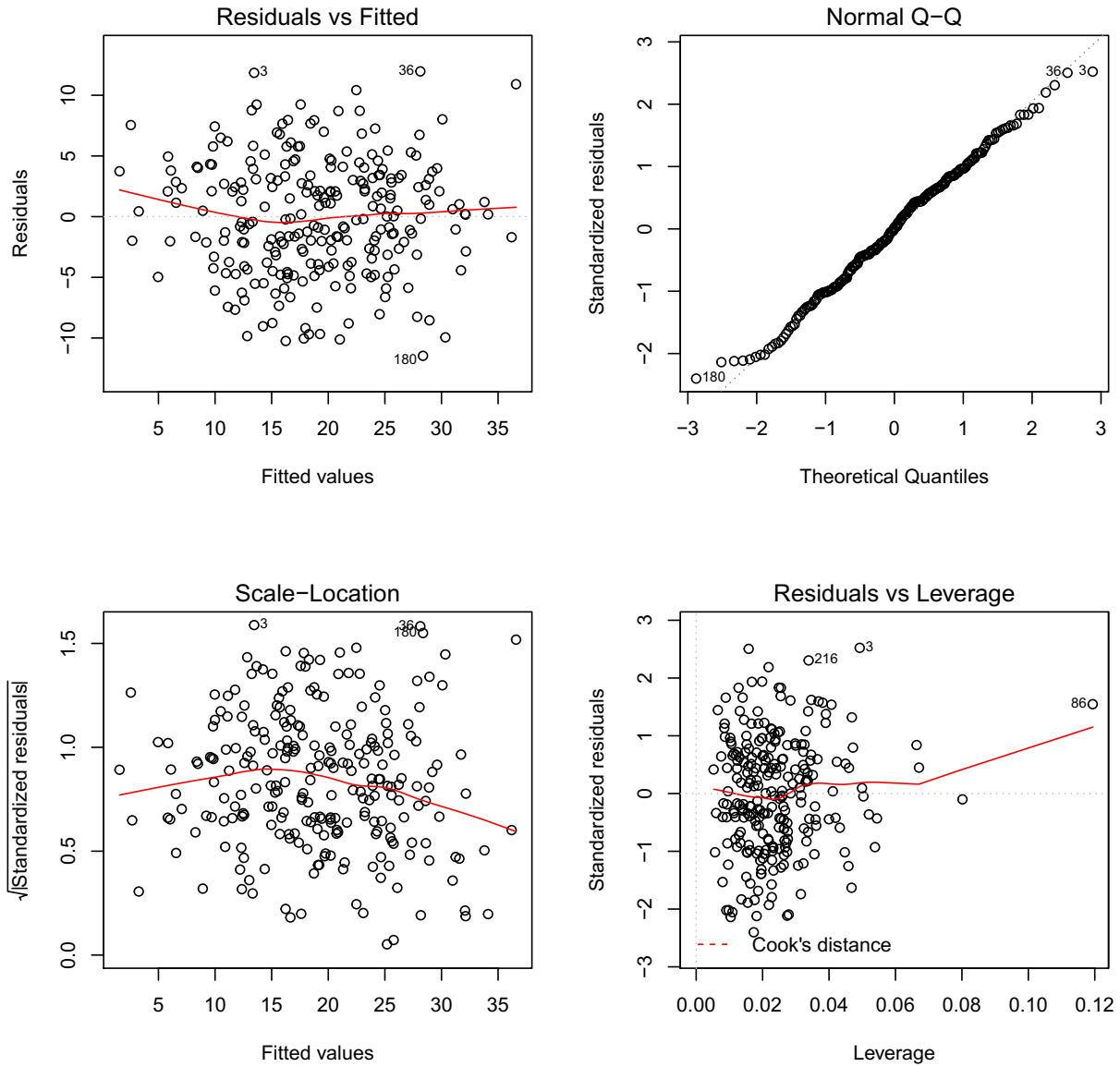
As first step, we again estimate a “global model” including all predictors:

```
##
## Call:
## lm(formula = formula2, data = bodyfat, x = T, y = T)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.467  -3.374   0.025   3.248  11.967
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -181.9706    10.0971  -18.022 < 2e-16 ***
## allometry   109.5334    19.3485   5.661 4.19e-08 ***
## beergut      71.7376     8.1552   8.797 2.57e-16 ***
## heavysset   107.6988    65.6019   1.642 0.101934
## fleshiness  18.3580     5.2098   3.524 0.000508 ***
## age_stand    1.6371     0.3337   4.906 1.70e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.816 on 245 degrees of freedom
## Multiple R-squared:  0.6719, Adjusted R-squared:  0.6652
## F-statistic: 100.4 on 5 and 245 DF,  p-value: < 2.2e-16
```

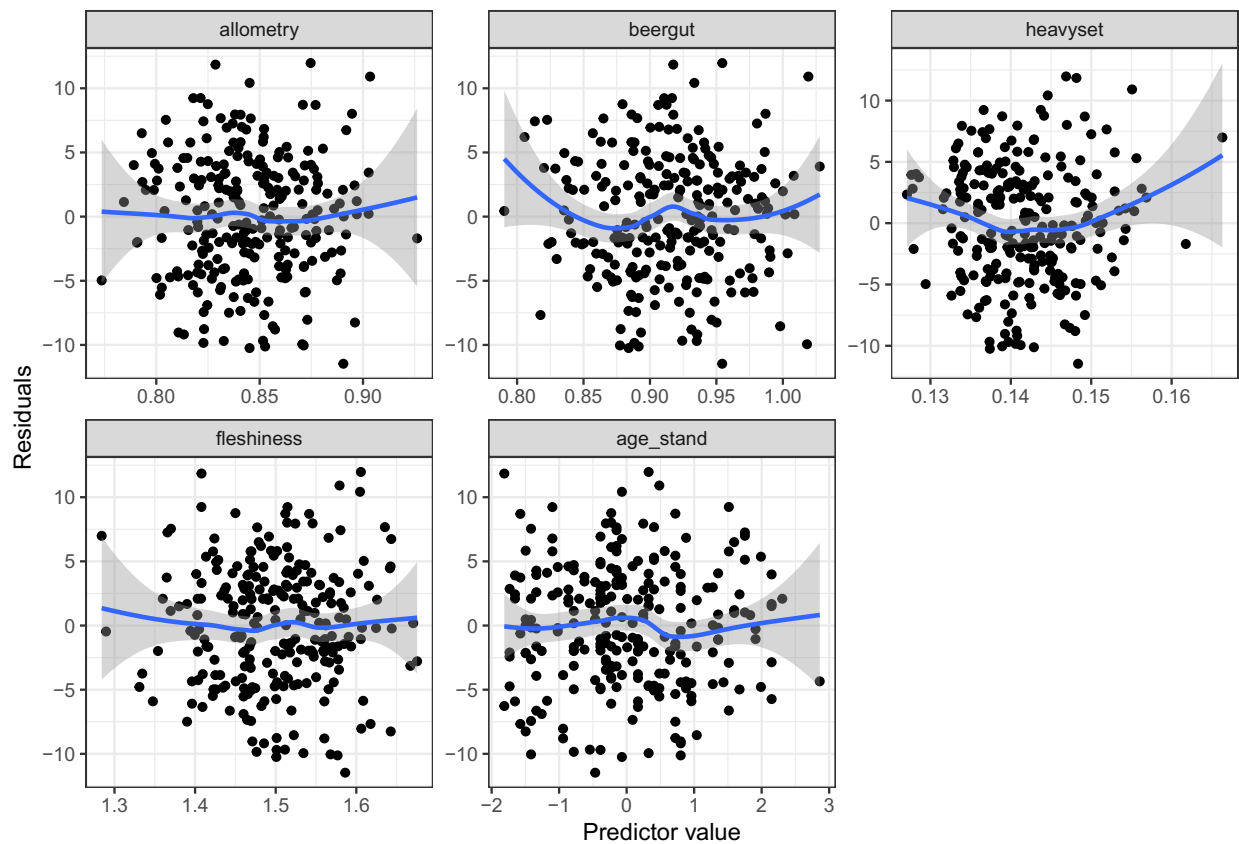


**Model diagnostics**

We again investigate the residuals of the model to detect possible misspecification:



Residuals plotted against predictor values suggest slight non-linear relations for beergut and heavysset since the pattern are a bit U-shaped:



Again, our approach to deal with this possible misspecification is to estimate robust standard errors, but to stick with the linearity assumption for the sake of interpretability:

```
##
## t test of coefficients:
##
##           Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -181.97060    9.90744 -18.3671 < 2.2e-16 ***
## allometry   109.53340   18.96208  5.7764 2.305e-08 ***
## beergut     71.73761    7.90310  9.0771 < 2.2e-16 ***
## heavysset   107.69881   65.62180  1.6412 0.1020380
## fleshiness  18.35804    4.82785  3.8025 0.0001809 ***
## age_stand   1.63708    0.33305  4.9154 1.623e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## BE selected DR model

Selection by backward elimination with AIC did not discard any variables because the highest p-value in the global model (0.102) is smaller than our significance level of 0.157 corresponding to AIC:

```
##  
## Call:  
## lm(formula = formula2, data = bodyfat, x = T, y = T)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max  
## -11.467  -3.374   0.025   3.248  11.967  
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -181.9706    10.0971  -18.022 < 2e-16 ***  
## allometry   109.5334    19.3485   5.661 4.19e-08 ***  
## beergut     71.7376     8.1552   8.797 2.57e-16 ***  
## heavyset   107.6988    65.6019   1.642 0.101934  
## fleshiness  18.3580     5.2098   3.524 0.000508 ***  
## age_stand   1.6371     0.3337   4.906 1.70e-06 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 4.816 on 245 degrees of freedom  
## Multiple R-squared:  0.6719, Adjusted R-squared:  0.6652  
## F-statistic: 100.4 on 5 and 245 DF,  p-value: < 2.2e-16
```

The adjusted  $R^2$  in this alternative model (0.672) is lower than in the previous selected model considering all circumference measurements separately (0.741).

## Stability of BE selected DR model

While the backward elimination did not actually remove any predictor from the model, in principle it could have done so, so one should still explore the stability of that model.

### Summary of the model stability

Below we report the coefficients in the global, in the BE selected DR model, the coefficients based on their bootstrap distributions, the corresponding VIFs, RCBs and RMSD ratios.

Most of the new variables seem to be highly relevant. In particular, fleshiness and standardized age were selected in more than 90% of the models and allometry and beergut were always chosen. Interestingly, heavysset only achieved a VIF of 37%.

##	Coef_global	SE_rob_global	Coef_sel	SE_rob_sel	Median_b	Lower_b
## allometry	109.53	19.35	109.53	18.96	115.17	73.20
## beergut	71.74	8.16	71.74	7.90	71.62	56.66
## age_stand	1.64	0.33	1.64	0.33	1.64	0.96
## fleshiness	18.36	5.21	18.36	4.83	18.17	7.88
## heavysset	107.70	65.60	107.70	65.62	103.38	0.00
##	Upper_b	VIF	RCB	RMSD	ratio	
## allometry	153.40	1.00	3.95	1.13		
## beergut	86.80	1.00	0.02	0.97		
## age_stand	2.31	1.00	0.23	1.05		
## fleshiness	27.75	0.93	-0.57	1.04		
## heavysset	227.32	0.37	37.72	1.27		

### MSF

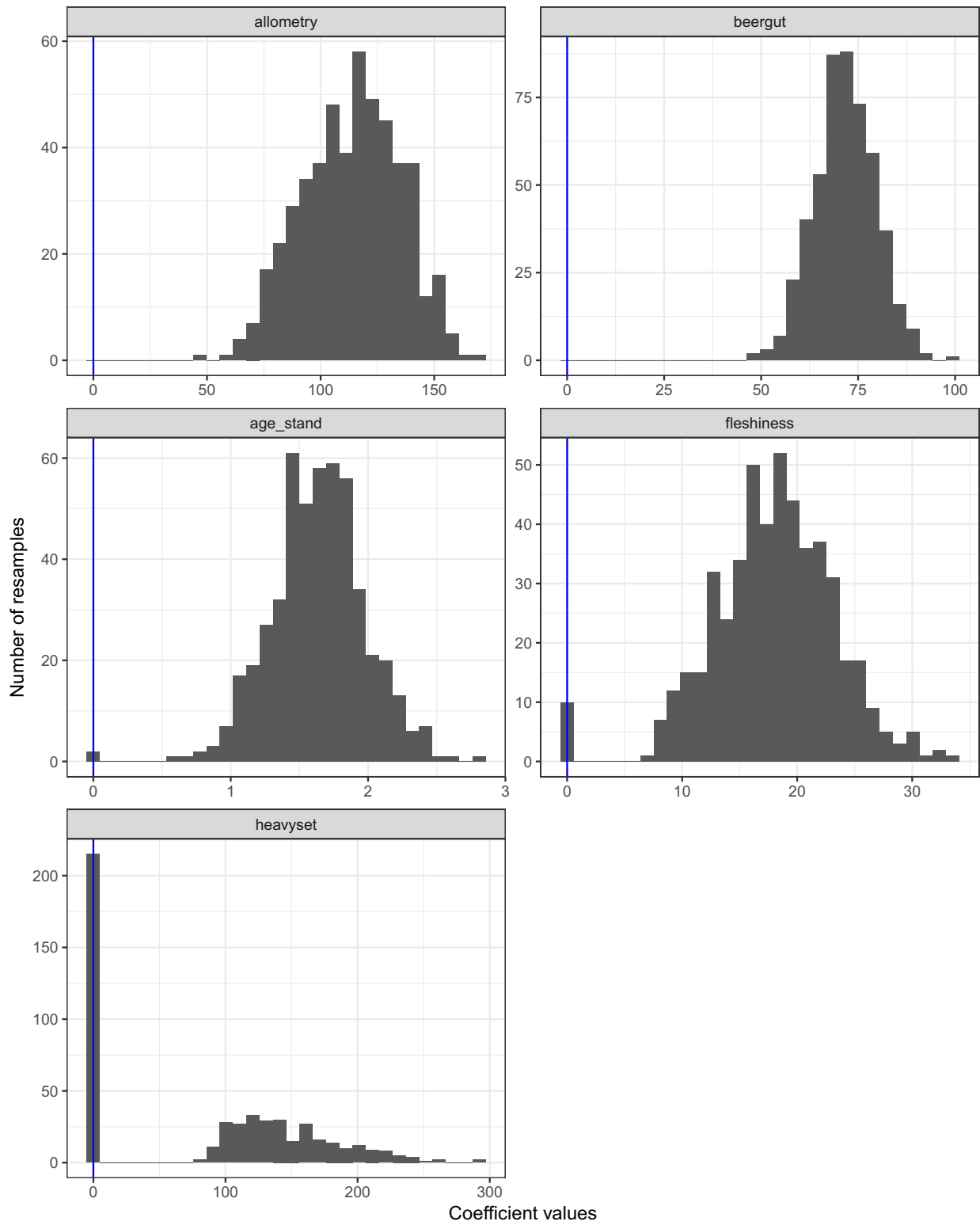
Here we report the list of BE selected DR models during the resampling procedure:

```
##
## (Intercept)+allometry+beergut+fleshiness+age_stand
## 0.560
## (Intercept)+allometry+beergut+heavysset+fleshiness+age_stand
## 0.364
## (Intercept)+allometry+beergut+age_stand
## 0.068
## (Intercept)+allometry+beergut+heavysset+age_stand
## 0.006
## (Intercept)+allometry+beergut+heavysset+fleshiness
## 0.002
```

The MSF are much higher now since fewer variables were considered and multicollinearity is not that strong. Our selected model ranks second with an MSF of 36.4%.

### Resampling distribution of predictors

In the resampling distributions, the spikes at zero are less pronounced than in the BE selected model, and the coefficients do no longer change their signs between the resampled models. Nevertheless, there are still uncertainties when variable selection is 'offered' to the model building process.



## Sensitivity analysis: fitting the ABE selected DR model

As sensitivity check, we fitted a model with ABE and  $\tau = 0.05$ .

Unsurprisingly, the ABE selected DR model is the same as the BE selected DR model:

```
##
## t test of coefficients:
##
##           Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -181.97060    9.90744 -18.3671 < 2.2e-16 ***
## allometry   109.53340   18.96208  5.7764 2.305e-08 ***
## beergut     71.73761    7.90310  9.0771 < 2.2e-16 ***
## heavysset   107.69881   65.62180  1.6412 0.1020380
## fleshiness  18.35804    4.82785  3.8025 0.0001809 ***
## age_stand   1.63708    0.33305  4.9154 1.623e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## Stability of ABE selected DR model

**Summary of the model stability** Below we report the coefficients from the global DR model, the ABE selected DR model, the bootstrap medians, 2.5th and 97.5th percentiles, and VIFs, RCBs and RMSD ratios.

All variables seem to be highly relevant. Fleshiness, standardized age, allometry and beergut were always chosen. When using ABE for selection variables, also heavysset is highly important for the estimation of bodyfat achieve a VIF of 96%.

```
##           Coef_global_sel SE_rob_global_sel Median_b Lower_b Upper_b VIF
## allometry           109.53           18.96  110.25  72.85  144.81 1.00
## beergut              71.74            7.90   71.74  56.89   86.97 1.00
## fleshiness           18.36            4.83   18.41   8.45   27.75 1.00
## age_stand            1.64            0.33    1.62   0.96    2.29 1.00
## heavysset           107.70           65.62  103.61 -20.62  227.32 0.96
##           RCB RMSD ratio
## allometry    0.57    0.99
## beergut      0.17    0.96
## fleshiness   0.15    0.97
## age_stand   -1.21    1.01
## heavysset    0.80    0.98
```

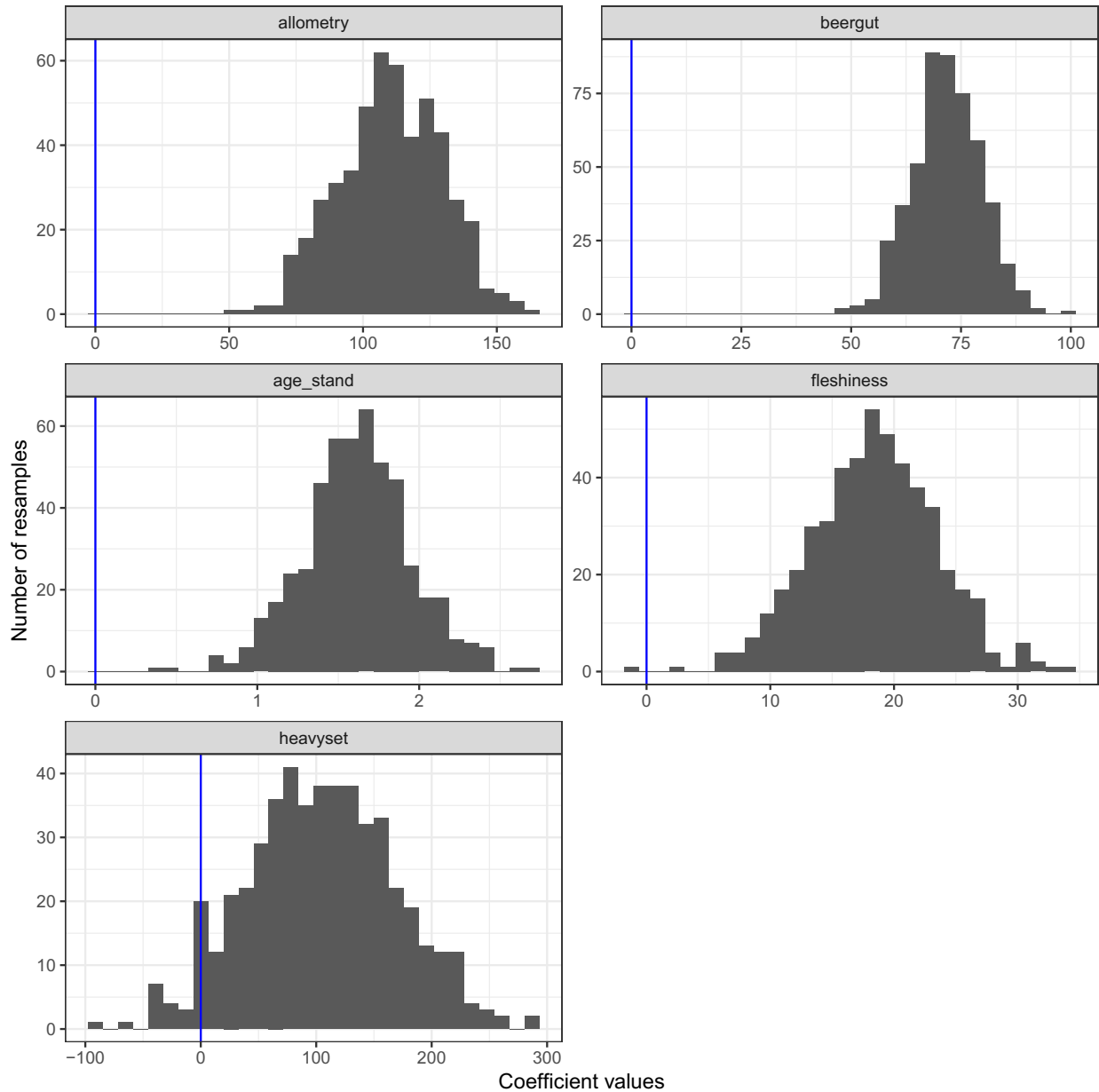
VIF, RCB and RMSD ratio clearly confirm the stability of the model with the alternative predictors.

**MSF** Here we report the list of selected models ranked by their MSF:

```
##
## (Intercept)+allometry+beergut+heavysset+fleshiness+age_stand
##                                     0.96
## (Intercept)+allometry+beergut+fleshiness+age_stand
##                                     0.04
```

Only two models appeared in the resampling procedure, where the global model with all predictors included dominates.

**Resampling distribution of predictors** For completeness, we also report the resampling distributions of the predictors. In only 4% of the models, heavyset was not selected, which leads to an only slightly elevated bar around the origin in its histogram.



## Conclusions

In this case study, we explored the instabilities incurred by variable selection by some resampling procedures and measures that were proposed in our paper (Heinze et al., 2018), further investigated in our follow-up study (Wallisch et al., 2021) and implemented in the R package *abe*.

We can conclude that the more stringent a variable selection procedure is, the more uncertainties are incurred. These uncertainties add up to the standard errors of regression coefficients, but are ignored in standard output of statistical software.

In some problems there may be a ‘sweet spot’ where the efficiency gain by removing irrelevant predictors outweighs the additional uncertainty incurred by offering selection to the estimation procedure. Clearly, such a sweet spot will be more likely to exist if:

- the sample size is large,
- if the candidate predictors have low correlation,
- and if the candidate predictors either have a strong association with the outcome or no or just irrelevant association with the outcome such that the variable selection algorithm can separate the true predictors from the non-predictors with high probability.

Users may use the code in this case study to investigate how more stringent criteria for variable selection affect uncertainties.

In the example, we also demonstrated how to increase stability of models by using derived variables that summarize several similar variables. Here we used domain expertise to derive these variables. Alternatively, one could apply explorative variable clustering techniques to detect such sets of variables that could be combined. There are some suggestions in that direction in the book of Harrell (2015).



## References

- Burnham KP, Anderson DA (2002). Model selection and multimodel inference. Springer.
- Harrell FE (2015). Regression Modeling Strategies, 2nd edition. Springer.
- Heinze G, Wallisch C, Dunkler D (2018). Variable selection - a review and recommendations for the practicing statistician. Biometrical Journal 60, 431-449, doi: 10.1002/bimj.201700067
- Johnson RW (1996). Fitting Percentage of Body Fat to Simple Body Measurements. Journal of Statistical Education 4,1. available at: <http://jse.amstat.org/v4n1/datasets.johnson.html>
- Wallisch C, Dunkler D, Rauch G, de Bin R, Heinze G (2021). Selection of variables for multivariable models: Opportunities and limitations in quantifying model stability by resampling. Statistics in Medicine 40, 369-381. DOI: 10.1002/sim.8779

## Session info

```
## R version 3.6.3 (2020-02-29)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 14393)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=German_Austria.1252 LC_CTYPE=German_Austria.1252
## [3] LC_MONETARY=German_Austria.1252 LC_NUMERIC=C
## [5] LC_TIME=German_Austria.1252
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets  methods   base
##
## other attached packages:
## [1] lmtest_0.9-38 zoo_1.8-8      sandwich_3.0-1  abe_3.0.1
## [5] corrplot_0.88 ggplot2_3.3.3 reshape2_1.4.4  vtable_1.3.1
## [9] kableExtra_1.3.4 knitr_1.31    mfp_1.5.2       survival_3.2-7
##
## loaded via a namespace (and not attached):
## [1] sjlabelled_1.1.8 tidyselect_1.1.0 xfun_0.23        purrr_0.3.4
## [5] splines_3.6.3     lattice_0.20-38  colorspace_2.0-0 vctrs_0.3.6
## [9] generics_0.1.0   htmltools_0.5.1.1 viridisLite_0.3.0 mgcv_1.8-31
## [13] yaml_2.2.1        utf8_1.1.4       rlang_0.4.10     pillar_1.6.0
## [17] glue_1.4.2        withr_2.4.1      DBI_1.1.1        lifecycle_1.0.0
## [21] plyr_1.8.6        stringr_1.4.0    munsell_0.5.0    gtable_0.3.0
## [25] rvest_1.0.0       codetools_0.2-16 evaluate_0.14     labeling_0.4.2
## [29] fansi_0.4.2      highr_0.8        Rcpp_1.0.6       scales_1.1.1
## [33] webshot_0.5.2    farver_2.0.3     systemfonts_1.0.1 digest_0.6.27
## [37] stringi_1.5.3    insight_0.14.0   dplyr_1.0.5      grid_3.6.3
## [41] tools_3.6.3      magrittr_2.0.1   tibble_3.0.6     crayon_1.4.1
## [45] pkgconfig_2.0.3  ellipsis_0.3.1   Matrix_1.2-18    xml2_1.3.2
## [49] assertthat_0.2.1 rmarkdown_2.7    svglite_2.0.0    httr_1.4.2
## [53] rstudioapi_0.13 R6_2.5.0         nlme_3.1-144     compiler_3.6.3
```

# Towards recommendations

## Recommendations

For our typical context (descriptive models, 10-25 candidate variables), we compiled some recommendations for the practicing statistician:

1. Generate initial set of variables
2. Decide on whether and which type of variable selection is needed
3. Perform stability investigations
4. Tackle post-selection inference
5. Reporting: recommendations for software developers

## 1. Generate initial set of variables

- Start with defensible assumptions on the roles of variables  
Use background knowledge to make assumptions on effect strength:  
strong or weak/unclear?
- Fit the global model

## 2. Decide whether and which variable selection is needed

- No selection applied to ,strong' variables
- No selection if number of candidate variables is small
- Selection for ,unclear' variables only with sufficient sample size
- If variable selection, do stability investigations
- See Table 3 (of Heinze et al 2018)

## 2. Decide whether and which variable selection is needed

- For descriptive and transparent prediction models, we prefer backward elimination with  $\alpha \geq 0.157$  starting with the global model
- $\alpha$  should be adjusted for sample size and purpose (larger  $\alpha$  in small samples, smaller  $\alpha$  in very large samples)
- If it should be guaranteed that important variables are not missed, augmented backward elimination can be recommended
- In small samples, perform penalized regression with a fixed penalty

## 2. Decide whether and which variable selection is needed

**TABLE 3** Some recommendations on variable selection, shrinkage, and stability investigations based on events-per-variable ratios

Situation	Recommendation
For some IVs it is known from previous studies that their effects are strong, for example age in cardiovascular risk studies or tumor stage at diagnosis in cancer studies.	Do not perform variable selection on IVs with known strong effects.
$EPV_{global} > 25$	Variable selection (on IVs with unclear effect size) should be accompanied by stability investigation.
$10 < EPV_{global} \leq 25$	Variable selection on IVs with unclear effect size should be accompanied by postestimation shrinkage methods (e.g. Dunkler et al., 2016), or penalized estimation (LASSO selection) should be performed. In any case, a stability investigation is recommended.
$EPV_{global} \leq 10$	Variable selection not recommended. Estimate the global model with shrinkage factor, or penalized likelihood (ridge regression). Interpretation of effects may become difficult because of biased effect estimation.

## 2. Decide whether and which variable selection is needed

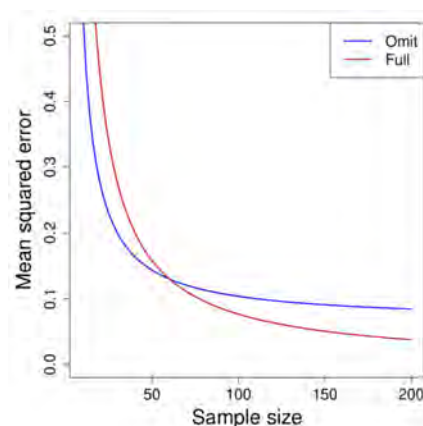
**TABLE 3** Some recommendations on variable selection, shrinkage, and stability investigations based on events-per-variable ratios

Situation	Recommendation
For some IVs it is known from previous studies that their effects are strong, for example age in cardiovascular risk studies or tumor stage at diagnosis in cancer studies.	Do not perform variable selection on IVs with known strong effects.
$EPV_{global} > 25$	Variable selection (on IVs with unclear effect size) should be accompanied by stability investigation.
$10 < EPV_{global} \leq 25$	Variable selection on IVs with unclear effect size should be accompanied by postestimation shrinkage methods (e.g. Dunkler et al., 2016), or penalized estimation (LASSO selection) should be performed. In any case, a stability investigation is recommended.
$EPV_{global} \leq 10$	Variable selection not recommended. Estimate the global model with shrinkage factor, or penalized likelihood (ridge regression). Interpretation of effects may become difficult because of biased effect estimation.

For definition and role of EPV in binary models, see Van Smeden et al, BMC MedResMeth 2016

## 2. Decide whether and which variable selection is needed

- In theory MSE of a regression coefficient could be lower if a companion covariate is omitted – if sample size is small
- But for practical causal inference, we could not find any relevant advantage of using backward elimination (as compared to using a global model with pre-selected confounders):
  - Luijken K, Groenwold R, van Smeden M, Strohmaier S, Heinze G, submitted: ‚Causal inference when selection of confounders is partly based on backward elimination: likely biased, rarely more efficient‘



### 3. Perform stability investigations

- Mandatory:
  - Computation of variable inclusion frequencies
  - Estimate sampling distribution of regression coefficients (bootstrap)
- Optional:
  - Model selection frequencies
  - Pairwise selection frequencies
  - Impact of variable selection on bias and variance  
We propose the bootstrap-based measures  
,Relative bias conditional on selection' (RBCS) and  
,Root mean squared difference ratio' (RMSDR) (see also Wallisch et al, 2021)
- Sensitivity analyses: change impact of decisions that were made

### 4. Tackle post-selection inference

Post-selection inference:

- Not much of a problem with large samples, say  $EPV > 100$  (BE  $\rightarrow$  BIC selection)
- Pragmatic solutions for smaller samples:
  - (i) *Situation:* The effect of an IV should be formally tested, but no theory exists on which subset of variables should be included in the model.  
*Solution:* Perform inference in the global model.
  - (ii) *Situation:* There exists a strong theory supporting only a small number of competing models.  
*Solution:* Perform multimodel inference with AIC.
  - (iii) *Situation:* There is no strong theory that can be used for model-building, but the global model is implausible. Although in this case effects of IVs can no longer be formally tested, still some evidence on the variability of estimates is needed.  
*Solution:* Perform multi-model inference with the resampled distribution of regression coefficients.
- At the very least, use robust standard errors (model mis-specification)!

## 5. Reporting: recommendations for software developers

- As long as packages as SPSS, SAS, or R's step() function output the ,final model' without adjustment for the selection process, nothing will change!
- Software developers should include a set of stability measures in the output
- SAS's PROC GLMSELECT is a start, yet still not acceptable
- R package `abe` for extended output

## 5. Reporting: recommendations for software developers

TABLE 5 Body fat study: global model, model selected by backward elimination with a significance level of 0.157 (AIC selection), and some bootstrap-derived quantities useful for assessing model uncertainty

Predictors	Global model			Selected model						
	Estimate	Standard error	Bootstrap inclusion frequency (%)	Estimate	Standard error	RMSD ratio	Relative conditional bias (%)	Bootstrap median	Bootstrap 2.5th percentile	Bootstrap 97.5th percentile
(Intercept)	4.143	23.266	100 (fixed)	5.945	8.150	0.97		5.741	-49.064	50.429
height	-0.108	0.074	100 (fixed)	-0.130	0.047	1.02	+4.9	-0.116	-0.253	0.043
abdomen	0.897	0.091	100 (fixed)	0.875	0.065	1.05	-2.1	0.883	0.687	1.050
wrist	-1.838	0.529	97.6	-1.729	0.483	1.07	-1.6	-1.793	-2.789	-0.624
age	0.074	0.032	84.6	0.060	0.025	1.14	+4.2	0.069	0	0.130
neck	-0.398	0.234	62.9	-0.330	0.219	1.24	+30.3	-0.387	-0.825	0
forearm	0.276	0.206	54.0	0.365	0.192	1.14	+46.6	0.264	0	0.641
chest	-0.127	0.108	50.9	-0.135	0.088	1.14	+68.0	-0.055	-0.342	0
thigh	0.173	0.146	47.9			1.13	+64.4	0	0	0.471
biceps	0.175	0.170	43.1			1.15	+101.4	0	0	0.541
hip	-0.149	0.143	41.4			1.08	+85.3	0	-0.415	0
ankle	0.190	0.220	33.5			1.11	+82.2	0	-0.370	0.605
weight	-0.025	0.147	28.3			0.95	+272.3	0	-0.355	0.295
knee	-0.038	0.244	17.8			0.78	+113.0	0	-0.505	0.436

RMSD, root mean squared difference, see Section 3.2(iv).



## Last minute: two further resources to consider

DE GRUYTER Int. J. Biostat. 2021; aop

Zihang Lu\* and Wendy Lou

### Bayesian approaches to variable selection: a comparative study from practical perspectives

<https://doi.org/10.1515/ijb-2020-0130>  
Received February 11, 2020; accepted February 27, 2021; published online March 24, 2021

**Abstract:** In many clinical studies, researchers are interested in parsimonious models that simultaneously achieve consistent variable selection and optimal prediction. The resulting parsimonious models will facilitate meaningful biological interpretation and scientific findings. Variable selection via Bayesian inference has been receiving significant advancement in recent years. Despite its increasing popularity, there is limited practical guidance for implementing these Bayesian approaches and evaluating their comparative performance in clinical datasets. In this paper, we review several commonly used Bayesian approaches to variable selection, with emphasis on application and implementation through R software. These approaches can be roughly categorized into four classes: namely the Bayesian model selection, spike-and-slab priors, shrinkage priors, and the hybrid of both. To evaluate their variable selection performance under various scenarios, we compare these four classes of approaches using real and simulated datasets. These results provide practical guidance to researchers who are interested in applying Bayesian approaches for the purpose of variable selection.

**Keywords:** Bayesian methods; linear regression; shrinkage; spike and slab; variable selection.

Projection predictive variable selection – A review and recommendations for the practicing statistician

Aki Vehtari  
First version 2018-03-06. Last modified 2021-05-12.

**1 Setup**

Load packages

```
library(mcmc)
library(rstanarm)
library(mcmc)
library(ggplot2)
library(parsnip)
library(caret)
library(MASS)
library(MASS)
library(MASS)
```

**2 Introduction**

This notebook was inspired by the article Hoerl, Wallisch, and Dunkler (2019), Variable selection – A review and recommendations for the practicing statistician. They provide an overview of various available variable selection methods that are based on significance or information criteria, penalized likelihood, the change-in-estimate criterion, background knowledge, or combinations thereof. I agree that they provide sensible recommendations and warnings for those methods. Similar recommendations and warnings hold for information criterion and naive cross-validation based variable selection in Bayesian framework as demonstrated by Penrose and Vehtari (2017a).

Inspired by our review, both reanalyzed the bodyfat data set

MEDICAL UNIVERSITY OF VIENNA

STRATOS INITIATIVE

Georg Heinze, Christine Wallisch, Daniela Dunkler  
CeMSIIS - Section for Clinical Biometrics

Part II-3  
16

## Conclusions

- Variable selection – loved by applied researchers, hated by statisticians
  - Their widespread use is our own fault!
  - Asymptotically, some VS methods work quite well
  - But: asymptotic properties – useless with real data?
  - We criticize ‚wrong‘ CIs with VS, but other methods (ML) don't even supply them (?)
- Software makes applications
  - SPSS, SAS, Stata, ...
  - Even R's `step()` just reports the last model, without comment
  - Stability investigations not reported in standard software (and standard applications)
  - Exception: SAS/PROC GLMSELECT, R package `abe`

## Conclusions

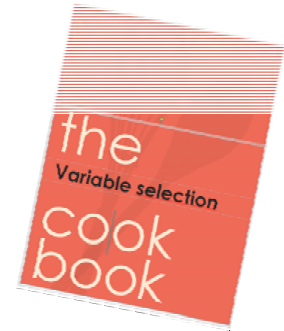
- We made some recommendations for quantities to be reported (by default in software!) if variable selection is applied (Implemented in Blagus' R package **abe**)
- We focus on models where regression coefficients should be interpretable
- Did not consider selection from a causal view (Witte & Didelez, BiomJ 2019)
- Did not cover nonlinearities and interactions
- Did not consider more fancy methods (SCAD, Alasso, Gradient boosting etc.)
  - (ECONOMETRICS-Paper)
  - Hard to believe that these methods would come without a cost?
  - Stability investigations still needed

## Outstanding issues – research required!

- From Sauerbrei et al, DAPR 2020:
  - ➔ 1. Investigation and comparison of the properties of **variable selection strategies**
  - ➔ 2. **Comparison of spline procedures** in 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?

## ***Recipe for disaster***

- Prepare a long list of poorly conceived predictors.
- Add only small  $n$ .
- Mix together in an extensive iterative data dredging.
- Select the model with the smallest  $p$ -values.
- Present this final model without further considerations.



*Bon appétit!*