Focus! SEM Trees and Forests for Identifying Moderators in Structural Equation Models

DAGStat 2025

Andreas M. Brandmaier

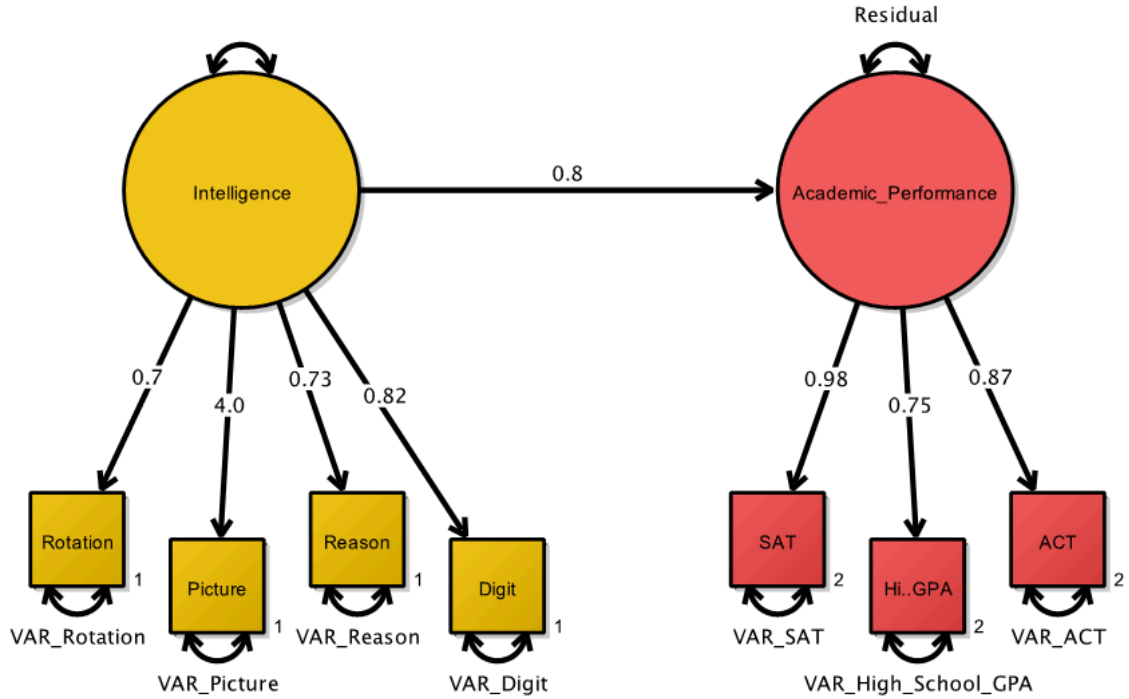
March 27, 2025

Question: “Given a (theory-based) multivariate model, which predictors/covariates/moderators are relevant?

SEM + Decision Trees + Focus Parameters + Random Forests + Variable Importance

Roadmap

SEM = Structure + Measurement



SEM

- More formally, we assume that we have l variables of which p are latent variables

- In RAM notation, we define:

  - a covariance matrix $\mathbf{S} \in \mathbb{R}^{l \times l}$ (“symmetric relations”),
  - a structural matrix $\mathbf{A} \in \mathbb{R}^{l \times l}$ (“asymmetric relations”),
  - and a filter matrix $\mathbf{F} \in \mathbb{R}^{p \times l}$ to filter out only observed variables

SEM

Then, the model-implied covariance matrix becomes:

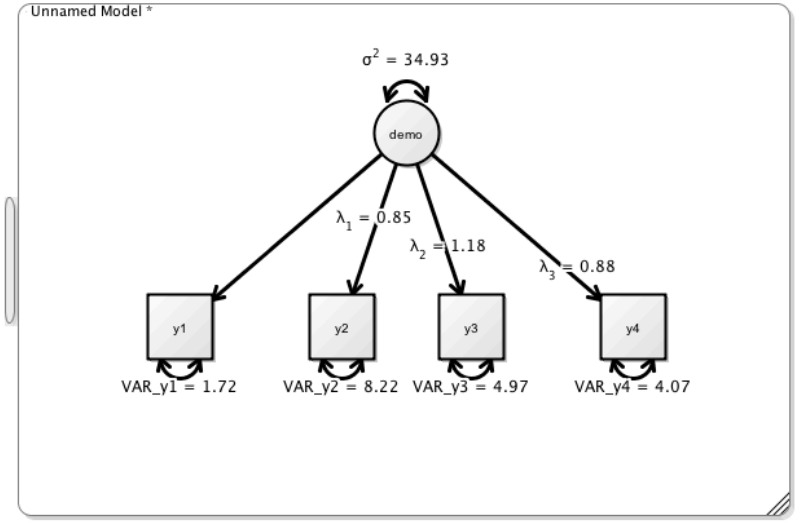
$$\mathbf{\Sigma} = \mathbf{F}(\mathbf{I_l} - \mathbf{A})^{-\mathbf{1}}\mathbf{S}(\mathbf{I_l} - \mathbf{A})^{-\mathbf{T}}\mathbf{F}^{\mathbf{T}}$$

And (covariance) likelihood fit function (based on multivariate normal assumption) of observed covariance S:

$$-2LL = ln|\mathbf{\Sigma}| + tr\left(\mathbf{\Sigma}^{-\mathbf{1}}\mathbf{S}\right) - ln\left|\mathbf{S}\right| - \mathbf{p}$$

(for multiple independent groups, the log-likelihoods sum up; means omitted here)

~~Commercial~~ Break: Ωnyx

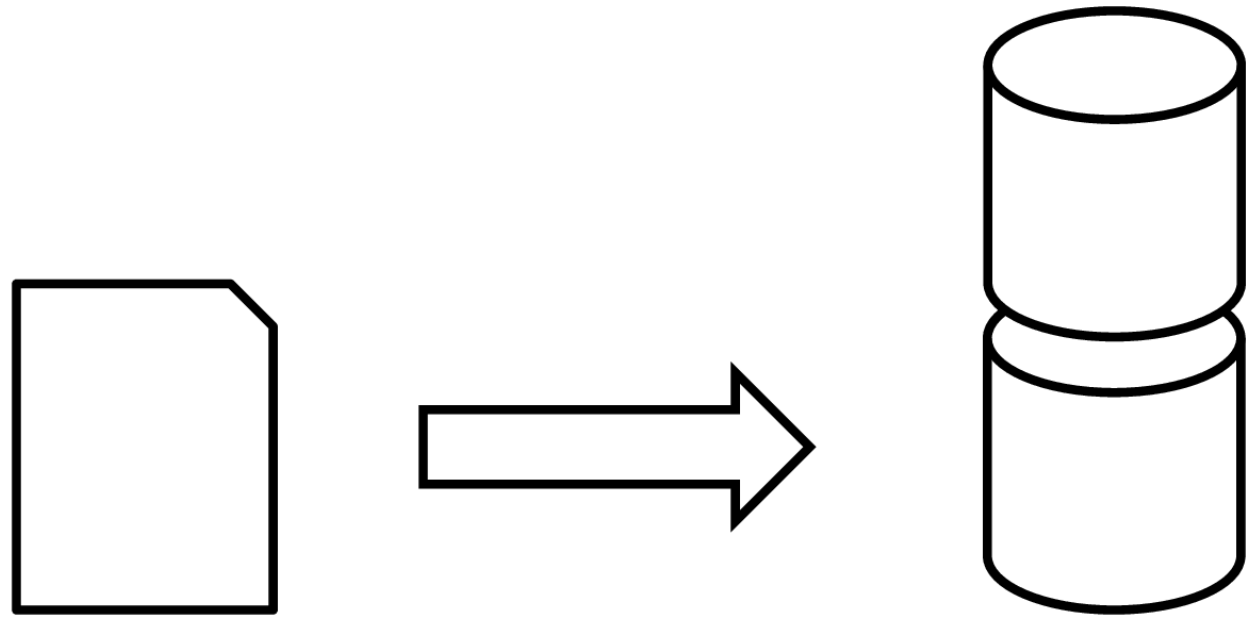


Lavaan code

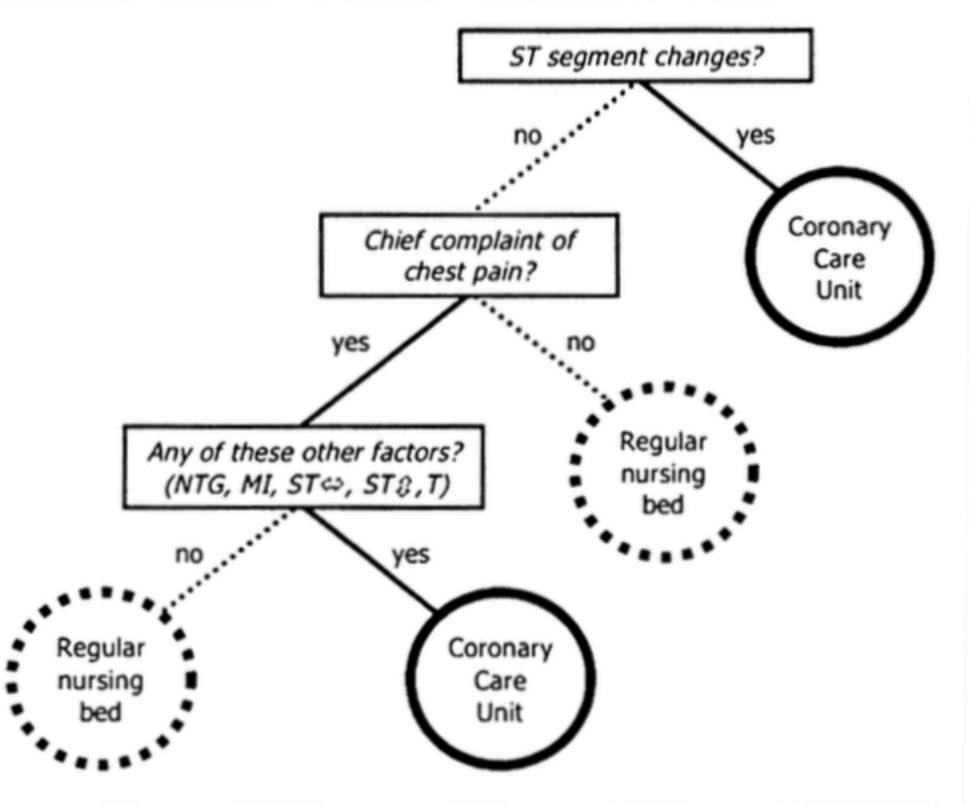
```
library(lavaan);
data<- ... ;
 onyx.model<-"
! regressions
   DEMO=~1.0*Y1
   DEMO=~Y2
   DEMO=~Y3
   DEMO=~Y4
! residuals, variances and covariances
   Y1 ~~ Y1
   Y2 ~~ Y2
   Y3 ~~ Y3
   Y4 ~~ Y4
   DEMO ~~ DEMO
! means
   Y1~0*1;
   Y2~0*1;
   Y3~0*1;
   Y4~0*1;
   DEMO~0*1;
";
fit<-lavaan(onyx.model, data=data);
summary(fit, fit.measures=TRUE);
```

https://onyx-sem.com/ and https://github.com/brandmaier/onyx and for Julia fans: https://github.com/StructuralEquationModels/StructuralEquationModels.jl

Theory-driven modeling

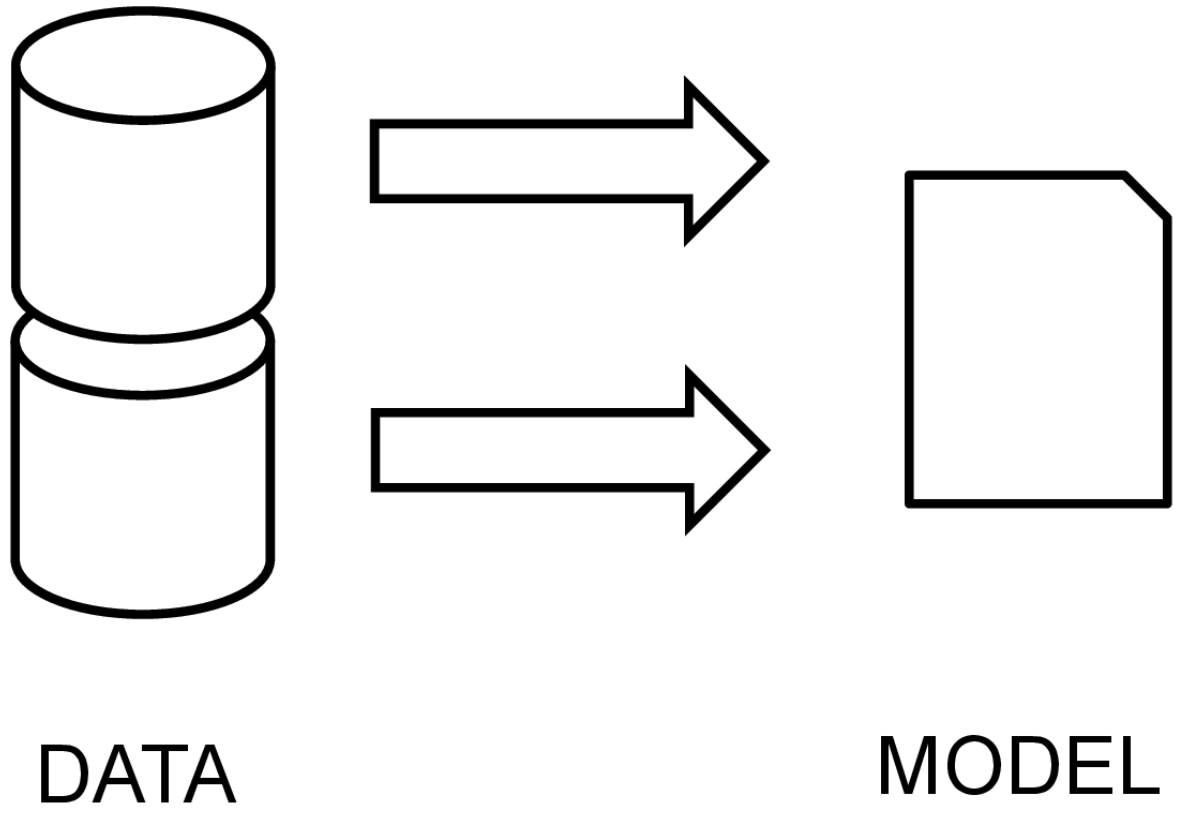


Decision Trees



Gigerenzer and Kurzenhaeuser (2005)

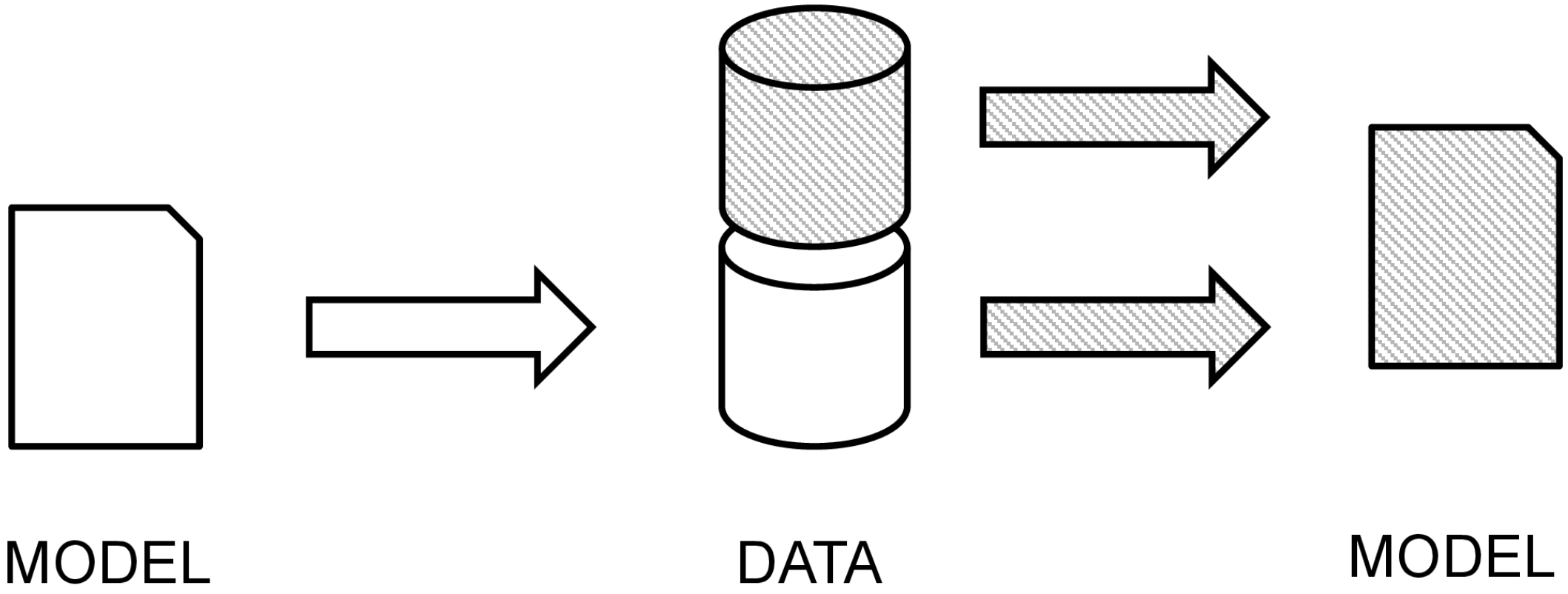
Data-driven modeling



SEM Trees

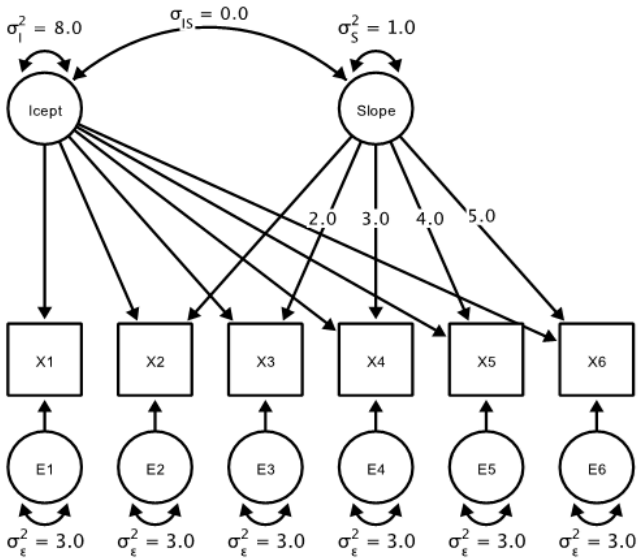
an instance of model-based recursive partitioning (Zeileis, Hothorn, and Hornik, 2008; Brandmaier, von Oertzen, McArdle, and Lindenberger, 2013)

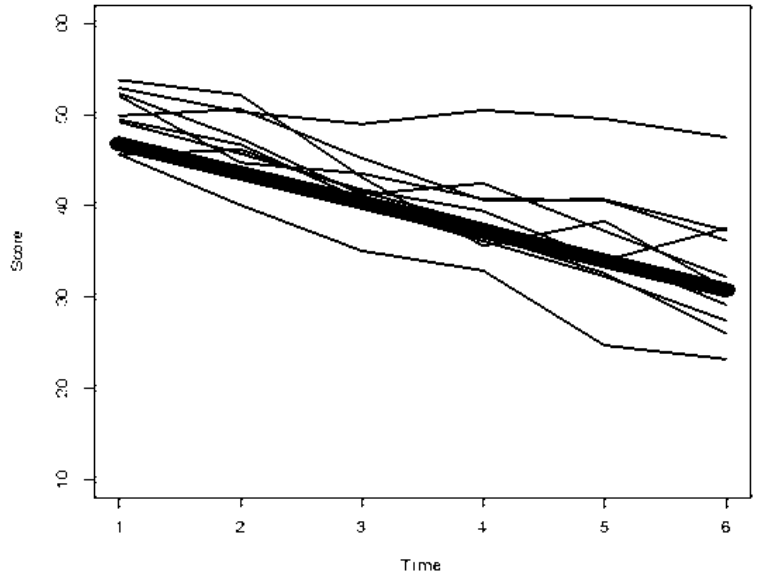
Theory-guided exploration



Brandmaier, Prindle, McArdle, and Lindenberger (2016)

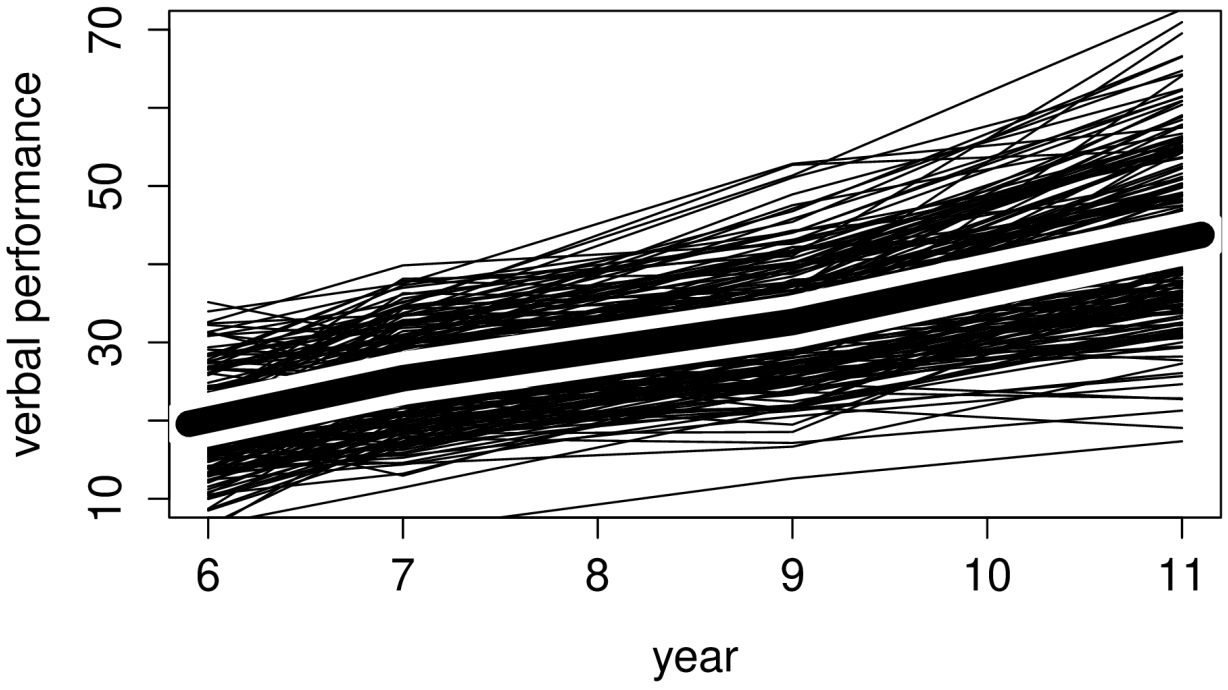
A Simple Example: Wechsler Intelligence Scale for Children





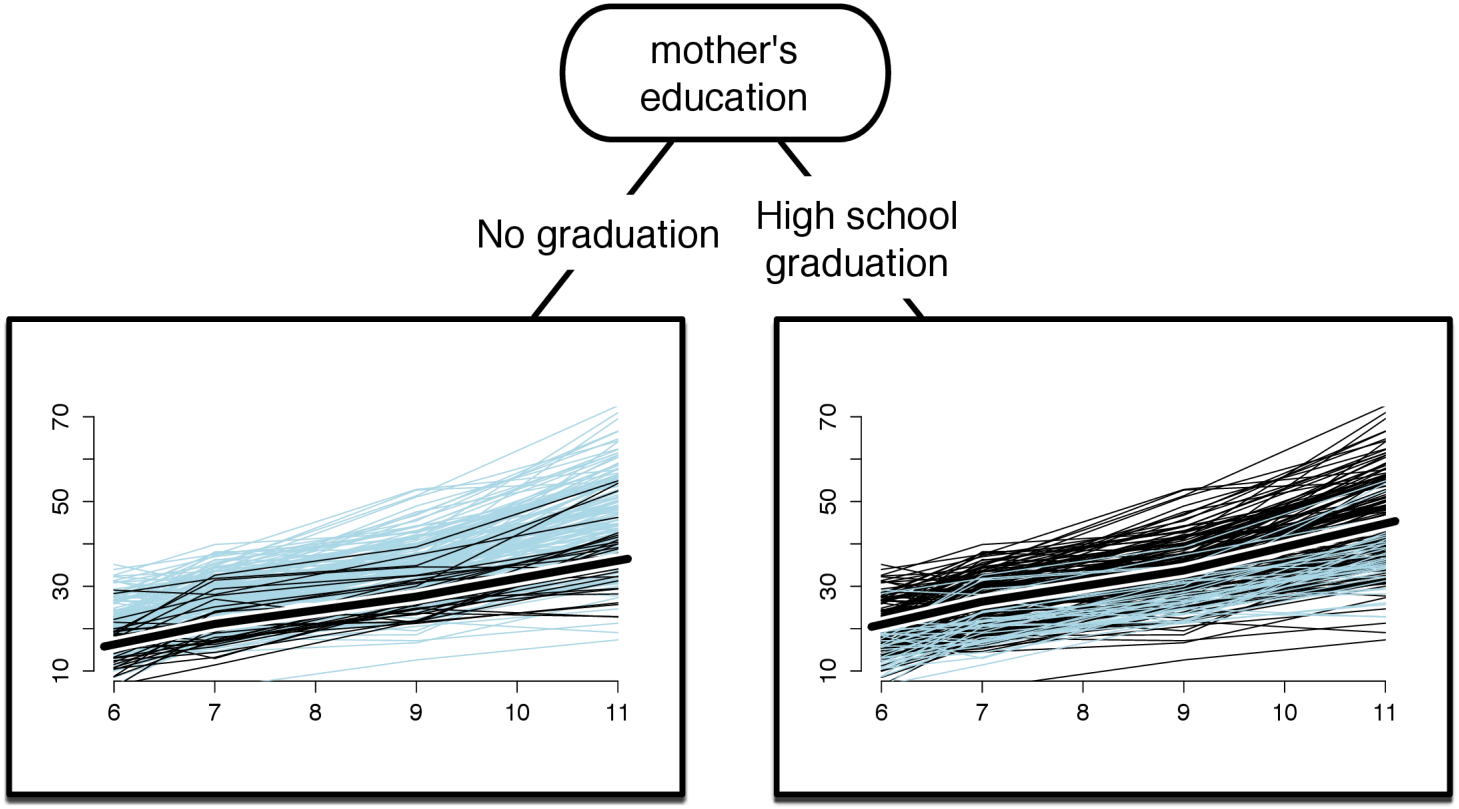
Brandmaier, von Oertzen, McArdle et al. (2013)

A Simple Example: WISC

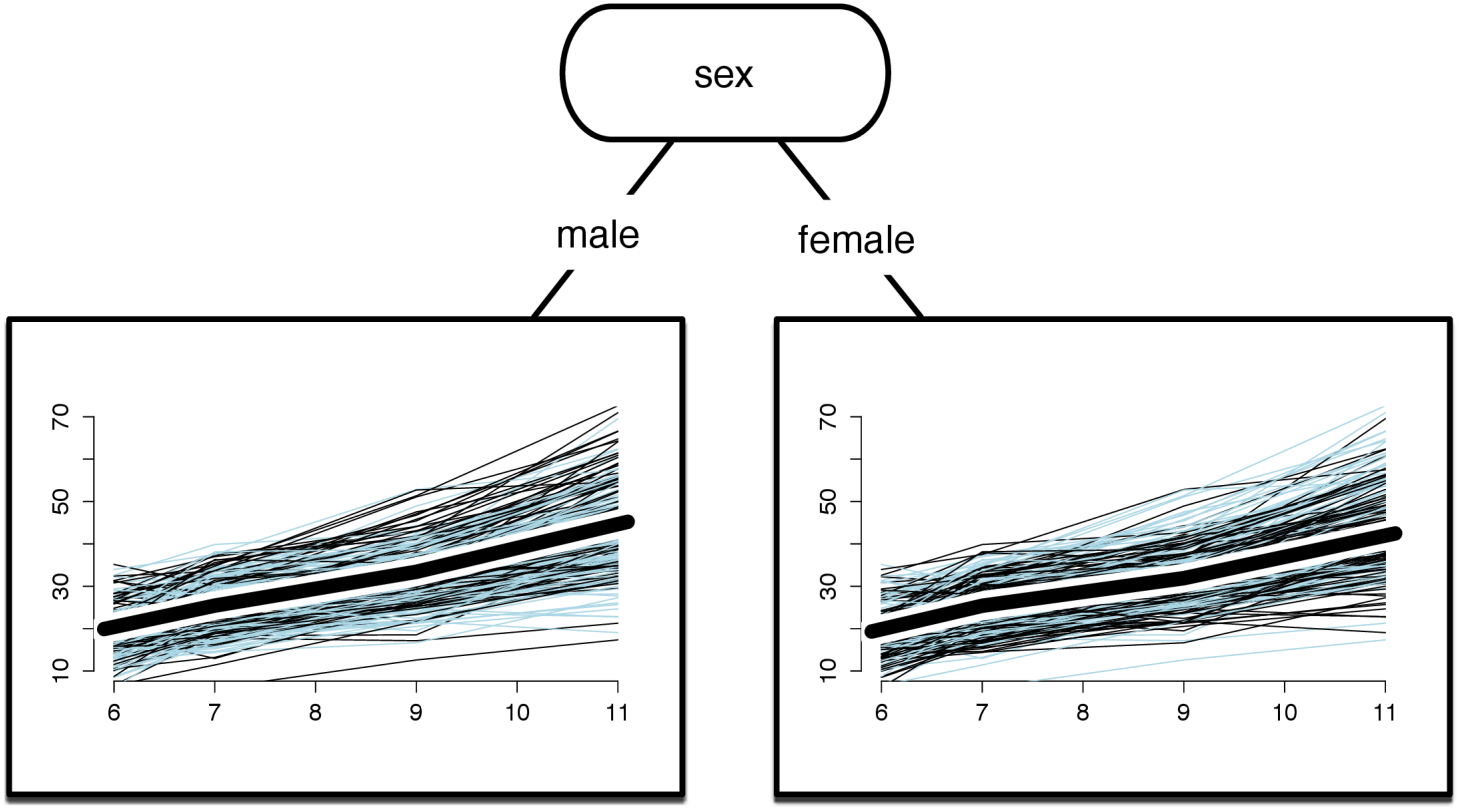


N=204 children, McArdle & Epstein, 1987

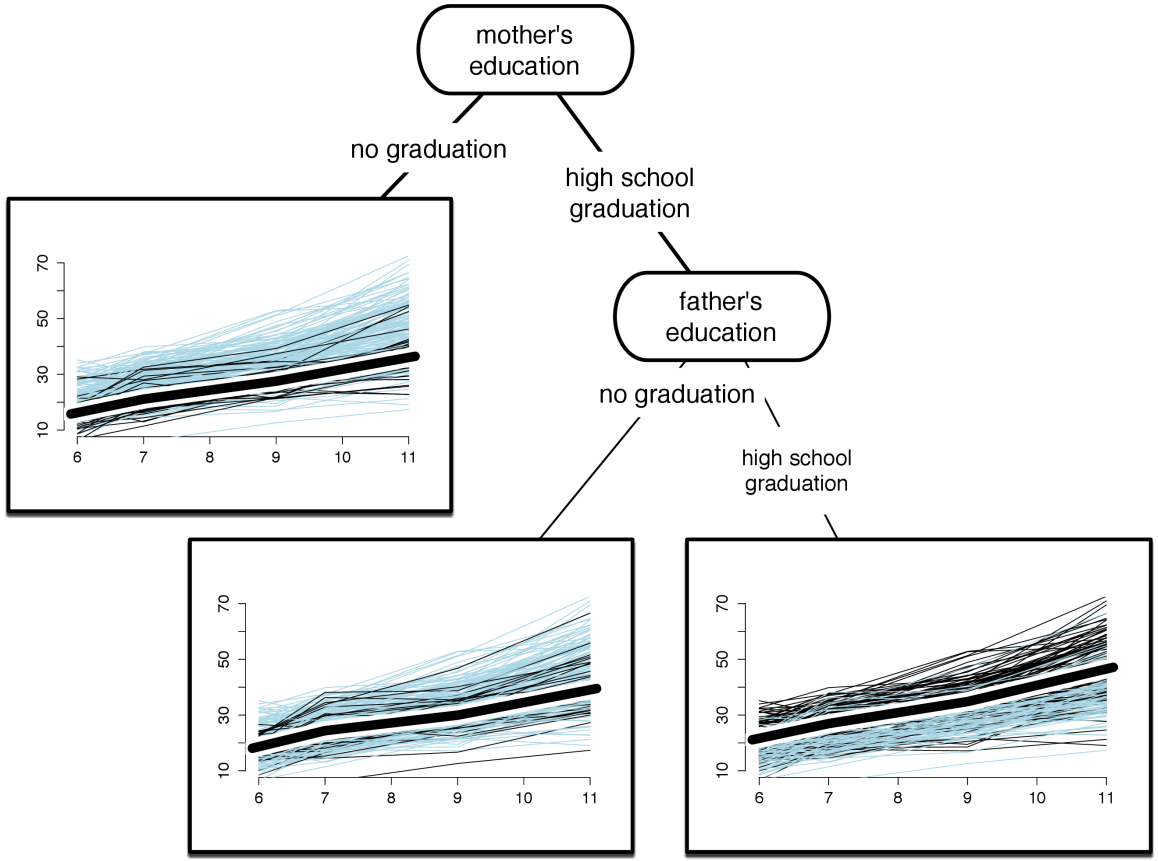
A Simple Example: WISC



A Simple Example: WISC



A Simple Example: WISC



What do splits represent?

Splits in an LGCM may represent any combination of:

- Differences in mean within-person changes
- Differences in interindividual differences in within-person change
- Differences in the mean of the intercept
- Differences in interindividual differences in the intercept
- Differences in the correlation of intercept and within-person change
- Differences in reliability or growth curve misfit (differences in measurement error as part of the model residual variance)

Evaluating Splits in (SEM) Trees

Testing for group differences

Brandmaier, von Oertzen, McArdle et al. (2013) proposed old-school likelihood ratio test for *split selection*:

Given a parametric model M with ML parameter estimates $\hat{\theta}$ and data X, which is exhaustively split into X_1 and X_2 with corresponding maximum likelihood estimates $\hat{\theta}_1$ and $\hat{\theta}_2$:

$$LR = -2LL\left(X_1|M(\hat{\theta}_1)\right) - 2LL\left(X_2|M(\hat{\theta}_2)\right) + 2LL\left(X|\hat{\theta})\right)$$

which is asymptotically χ^2-distributed if H_0 is true (i.e., no group differences) with $df = dim(\theta)$

Split selection

- Starting at the root of a tree, find best split by greedy search
- Depending on measurement scale
  - all possible split points for ordinal and metric variables (linear costs)
  - all possible dichotomizations for nominal variables (exponential costs)
- Continue splitting if difference is significant

Focus parameters with LR tests

For χ^2-based tests, instead of:

$$LR = -2LL\left(X_1|M(\hat{\theta}_1)\right) - 2LL\left(X_2|M(\hat{\theta}_2)\right) + 2LL\left(X|\hat{\theta})\right)$$

we estimate *loss of fit due to constraining only focus parameters* to identity across groups:

$$LR = -2LL\left(X_1 \mid M(\hat{\theta}_1)\right) - 2LL\left(X_2 \mid M(\hat{\theta}_2)\right) - 2LL\left(X_1 \mid M(\hat{\theta}'_1)\right) - 2LL\left(X_2 \mid M(\hat{\theta}'_2)\right)$$

with focus parameters constrained to be identical across groups for θ'_1 and θ'_2

Focus parameters

This is more expensive because we need to obtain maximum likelihood estimates for every possible split:

$$argmin_{\theta'_1,\theta'_2} -2LL\left(X_1|M(\theta'_1)\right) - 2LL(X_2|M(\theta'_2))\Big)$$

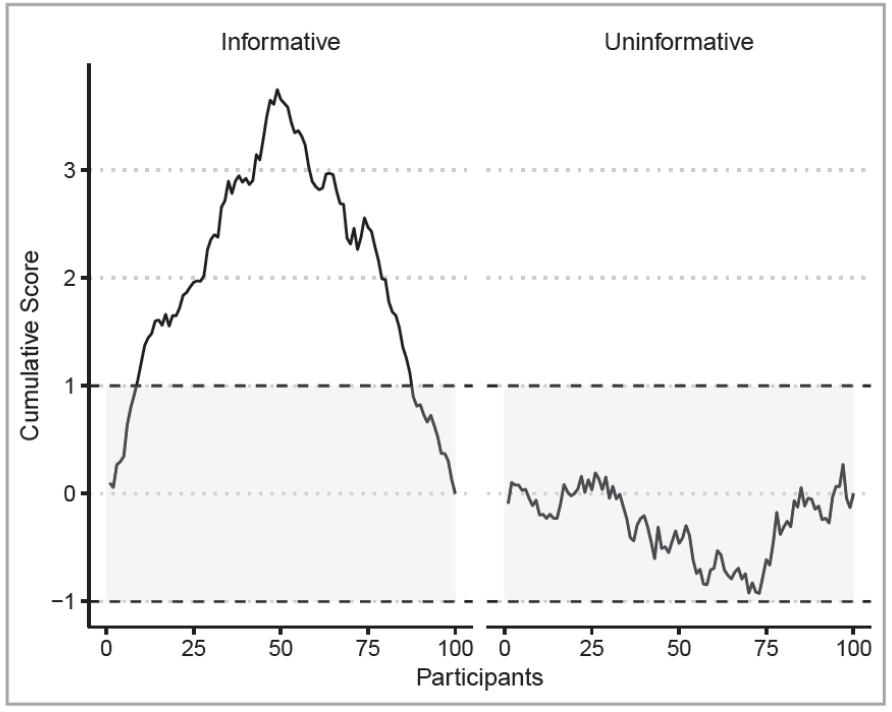
subject to focus parameters being identical

In total, we need one more iterative optimization process per each potential split point (with potentially more convergence issues)

Score-based Tests

- Likelihood ratio tests are computationally expensive, we get convergence problems, naive LR test suffer from multiple testing issues and variable selection bias like CART (Strobl, Boulesteix, Zeileis, and Hothorn, 2007)
- Let's use score tests (proposed for general model-based partitioning by Zeileis, Hothorn, and Hornik (2008)

Score-based Tests

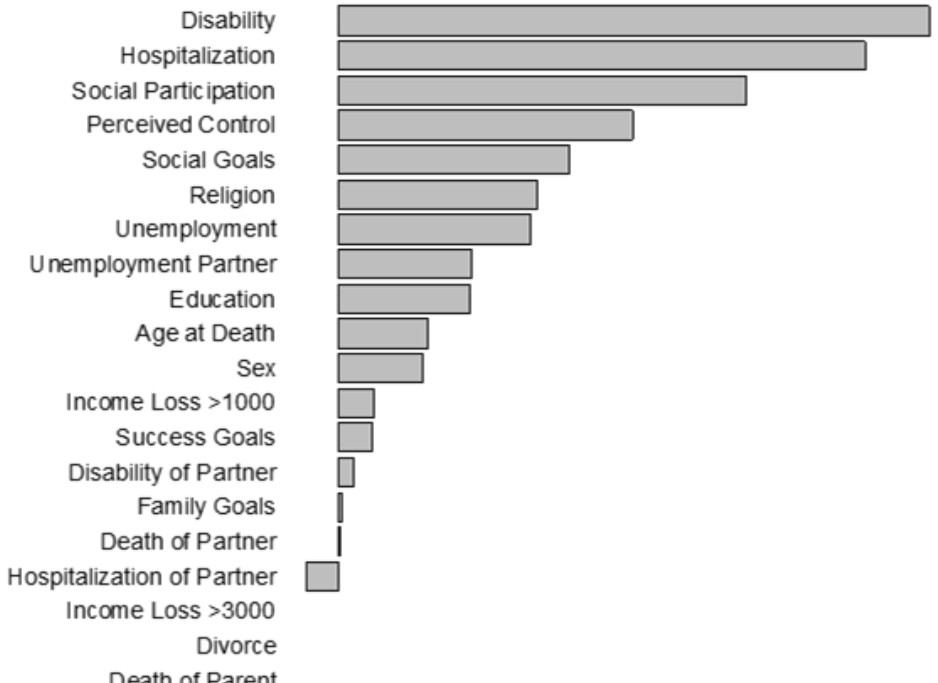


- Score based test statistics are functions of the case-wise derivatives of the (log)likelihood function
- Backbone of MOB in `party/partykit` and were also brought to SEM trees by (Arnold, Voelkle, and Brandmaier, 2021)
- Under H0 (no informative split), the cumulative score process is a Brownian bridge
- Various statistics available (e.g. maximum Lagrange multiplier $maxLM$; double maximum for illustration on the left)
- By computing them on subsets of parameters, we get *focus parameters* for free

Variable Importance

- Single trees are unstable
- Subsample data and predictors to create a forest with diverse predictor combinations
- Using a permutation approach, estimate contribution of each predictor to misfit (Brandmaier, Prindle, McArdle et al., 2016)
- ...but beware of marginal importance; (Strobl, Boulesteix, Kneib, Augustin, and Zeileis, 2008)

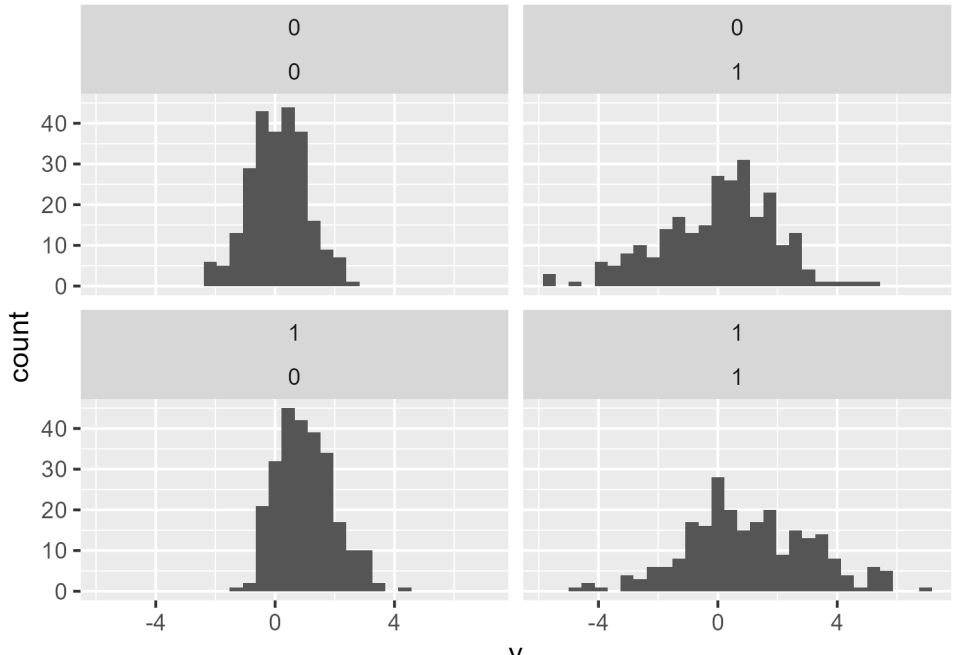
Example: Terminal decline of happiness from SOEP data (Brandmaier, Ram, Wagner, and Gerstorf, 2017)



A minimal example

Univariate predictions

Let's simulate some Gaussian data and two dichotomous predictors `pred_mean` and `pred_var` that perfectly predict differences in either location (0 vs 1) or scale (1 vs 2). Also, we throw three uninformative predictors in the mix (binomial distributed) and sample $1,000$ cases:



Run a tree

Specify model in `OpenMx` (or `lavaan`):

```
sem <- mxModel("Univariate Gaussian",
  type="RAM",  manifestVars="y",
               [...]
  # variance
  mxPath(from=manifests,arrows=2,free=TRUE,
     values = c(1), labels=c("var_y")),
# means
   mxPath(from="one",to=manifests, arrows=1,free=TRUE,
          values=c(0), labels=c("mean_y") )
```

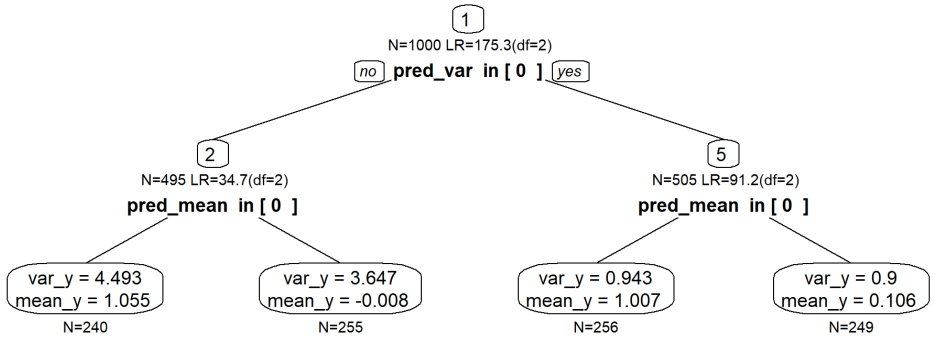
Run a tree

Run the tree

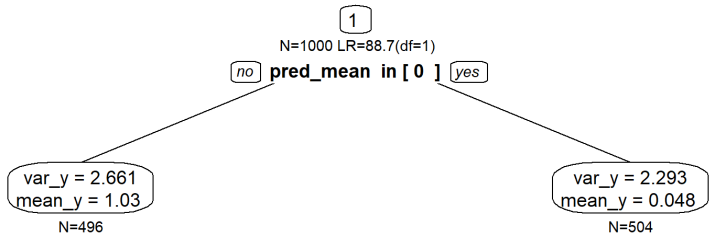
```
sem <- mxModel("Univariate Gaussian",
  type="RAM", manifestVars="y",
                [...]
  # variance
  mxPath(from=manifests,arrows=2,free=TRUE,
     values = c(1), labels=c("var_y")),
# means
   mxPath(from="one",to=manifests, arrows=1,free=TRUE,
          values=c(0), labels=c("mean_y") )

tree <- semtree(sem, simulated_data,
   control=semtree_control(method="score"),
   constraints=semtree.constraints(focus.parameters = "mean_y"))
```

Tree with and w/o focus

This seems to work...



(No focus)



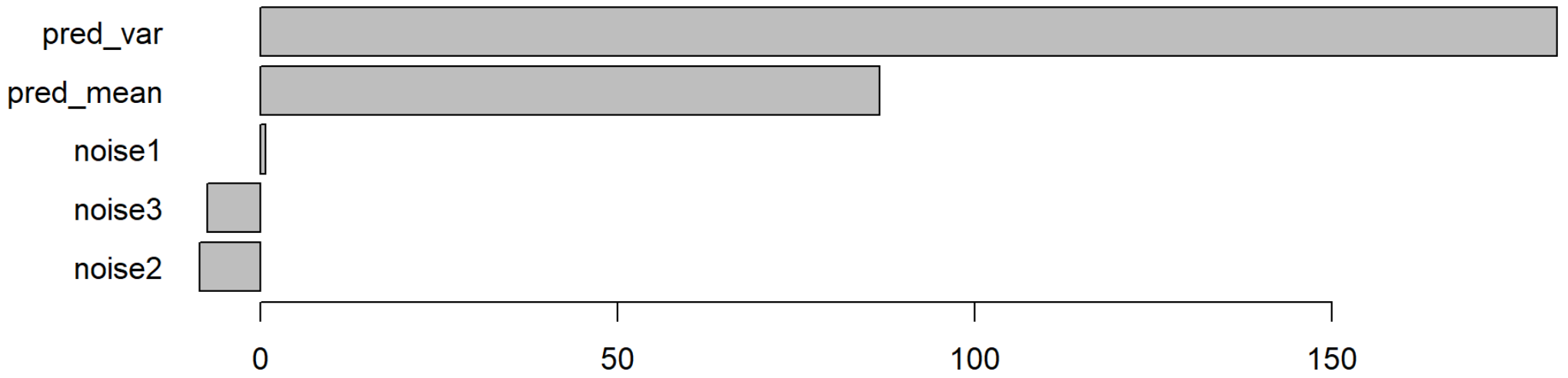
(focus on *mean_y*)

Run a forest

Run a SEM forest based on a single indicator model with two parameters (mean and variance) again *with focus parameter* and compute variable importance

```
forest <- semforest(sem, simulated_data,
      control=semforest_score_control(num.trees=100),
      constraints=semtree.constraints(focus.parameters = "mean_y"))

vim <- varimp(forest)
```

(Marginal) Variable importance estimate



This is flawed because the influence of *pred_var* should be about zero.

Problem

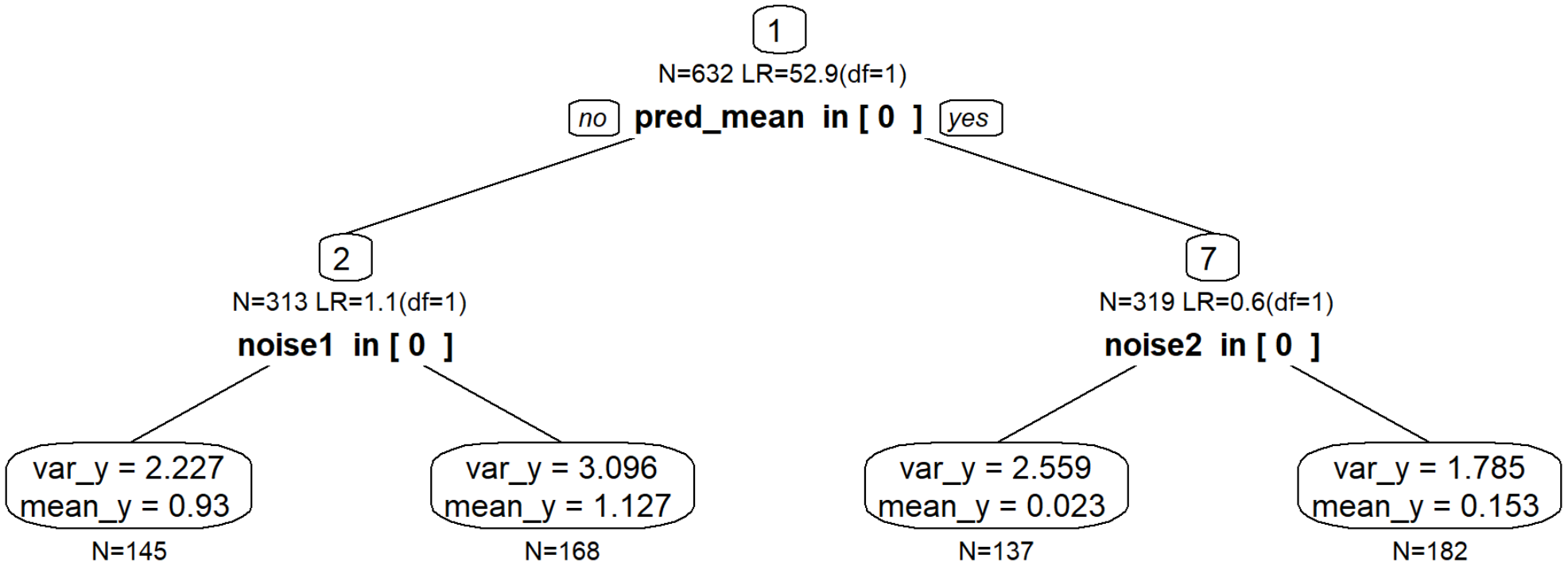
With SEM forest, we grow trees with

- subsampling of predictors (`mtry=2` or $\sqrt{m}$ or ..)
- no stopping rule ($\alpha = 1$) in order to explore deep conditional effects (~interactions) in the trees

Therefore, the permutation importance estimate will be influenced by differences with respect to all parameters of the model because all predictors will appear in the forest

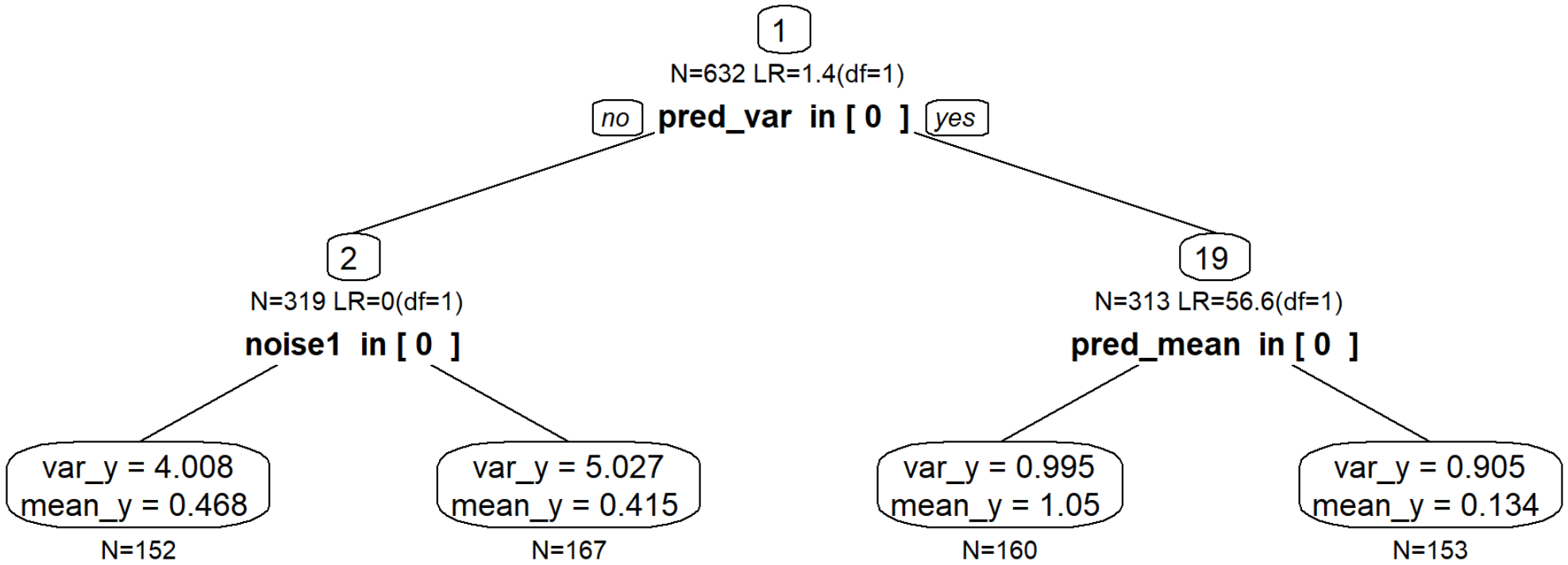
Inspect first tree

(Pruned) Tree from a forest with focus parameters



Inspect second tree

Another (pruned) Tree from a forest with focus parameters



Problem

Problem: This measure of variable importance considers differences w.r.t all parameters of the SEM

Solution: Estimate misfit incurred by only the focus parameters

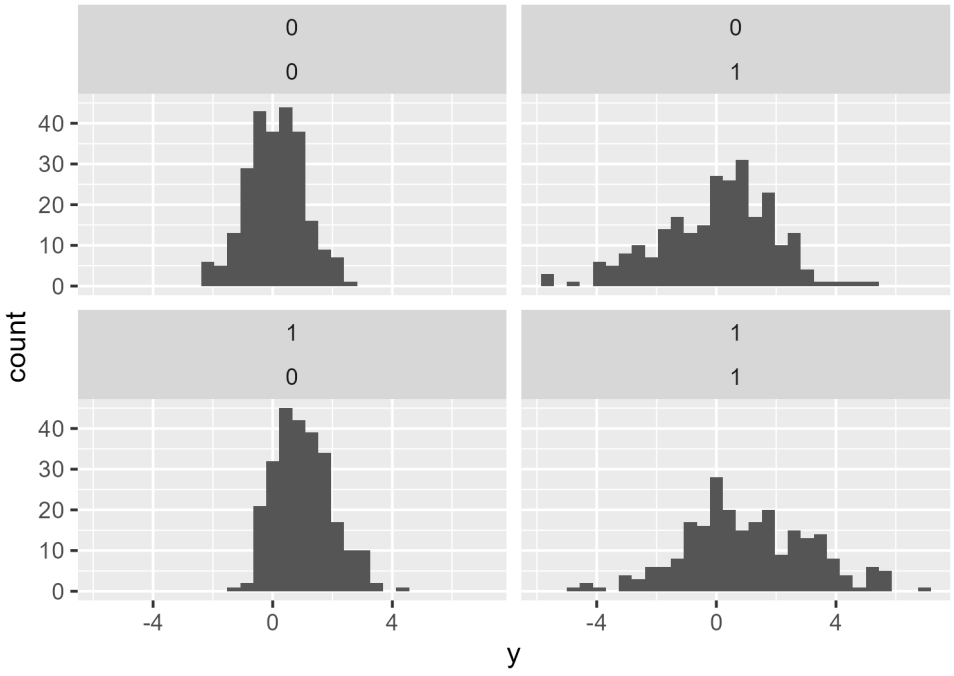
Algorithm (estimate importance for focus parameters)

Modified Importance Evaluation:

- For each tree, for each observation x, find leaf model M_1 by traversing the tree and leaf model M_2 by traversing the tree after permuting predictor in question
- Compute -2LL of x under M_1
- ~~Compute -2LL of x under M_2~~
- Compute -2LL of x under M_1 with only focus parameters plugged in from M_2
- Compute -2LL difference

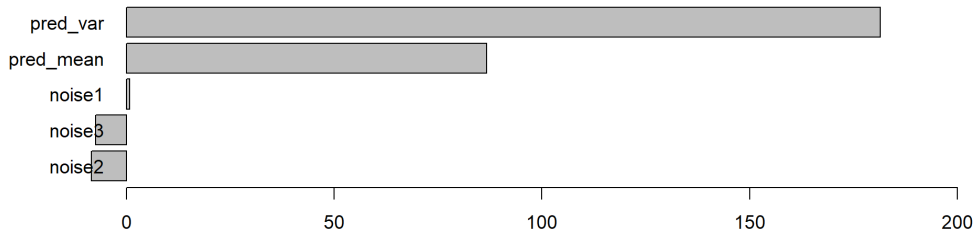
Back to the simulation

Simulated Data:

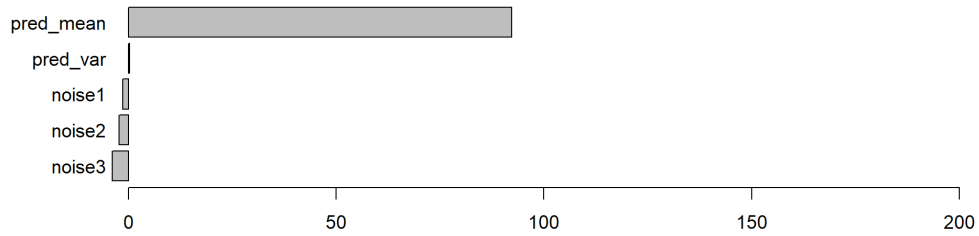


Importance Estimates

Forest with (score-based) focus parameter on mean:



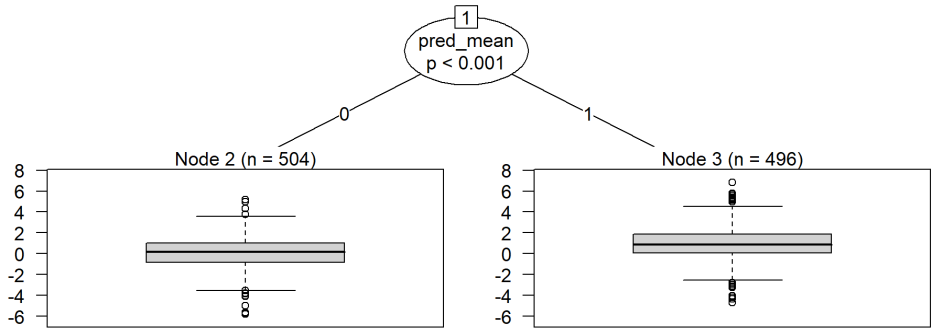
(old scheme)



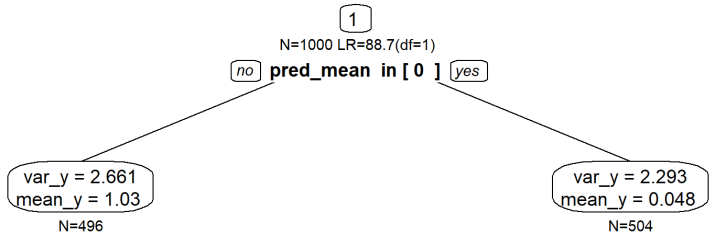
(new scheme)

To Focus or not to focus

Trees yield identical results in `partykit` and `semtree` (with focus parameter on mean):



made with ❤ by `partykit`



made with ❤ by `semtree`

W/o focus: potentially interesting because differences in variances may also be of interest (e.g., meaningful individual differences or reliability differences)

Some more examples

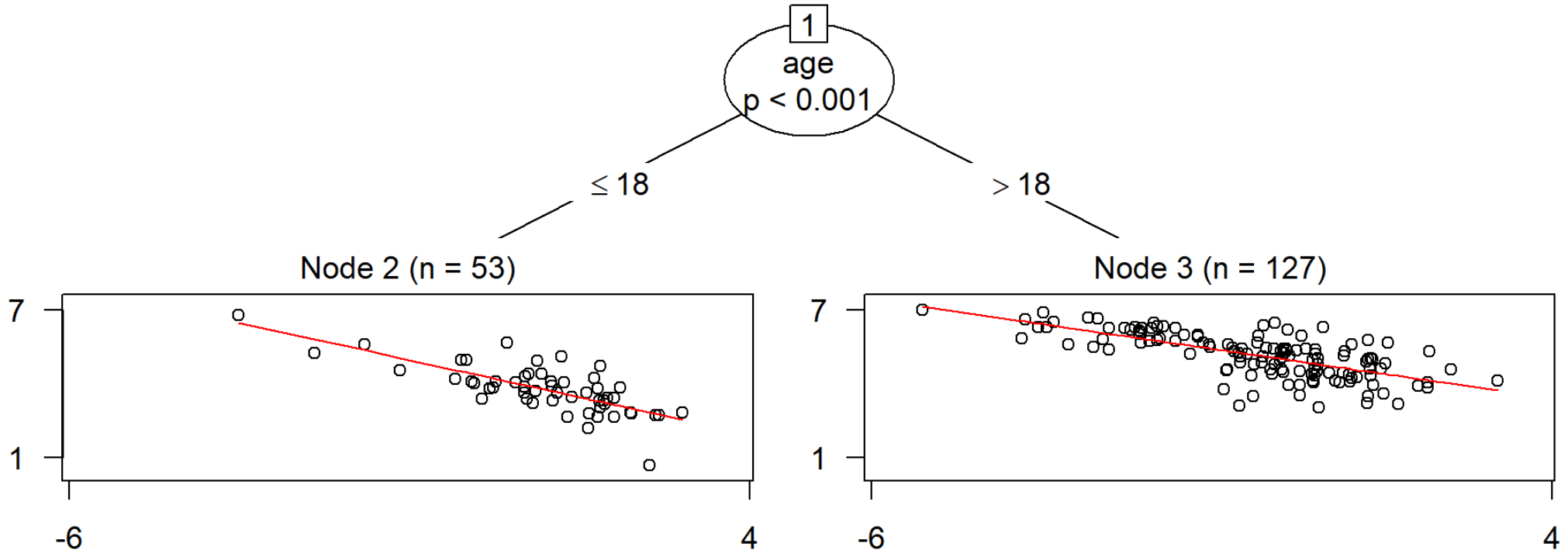
Journal Pricing

- The model to be partitioned is a linear regression for the number of library subscriptions by price per citation in log-log specification (i.e., with $k = 2$ coefficients) Zeileis, Hothorn, and Hornik (2008)

- Predictors: the raw price and number of citations, the age of the journal, number of characters and a factor indicating whether the journal is associated with a society or not.

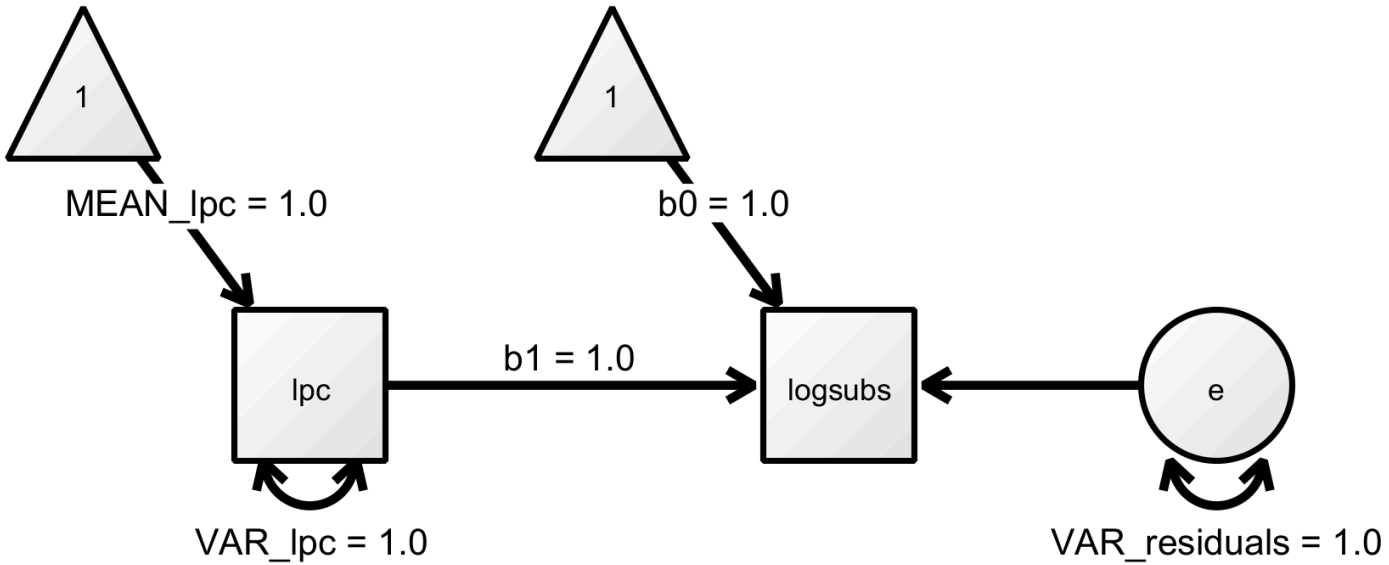
```
library(partykit)
j_tree <- partykit::lmtree(logsubs ~ lpc | price + citations +
                    + age + chars + society, data = Journals)
```

Journal Pricing



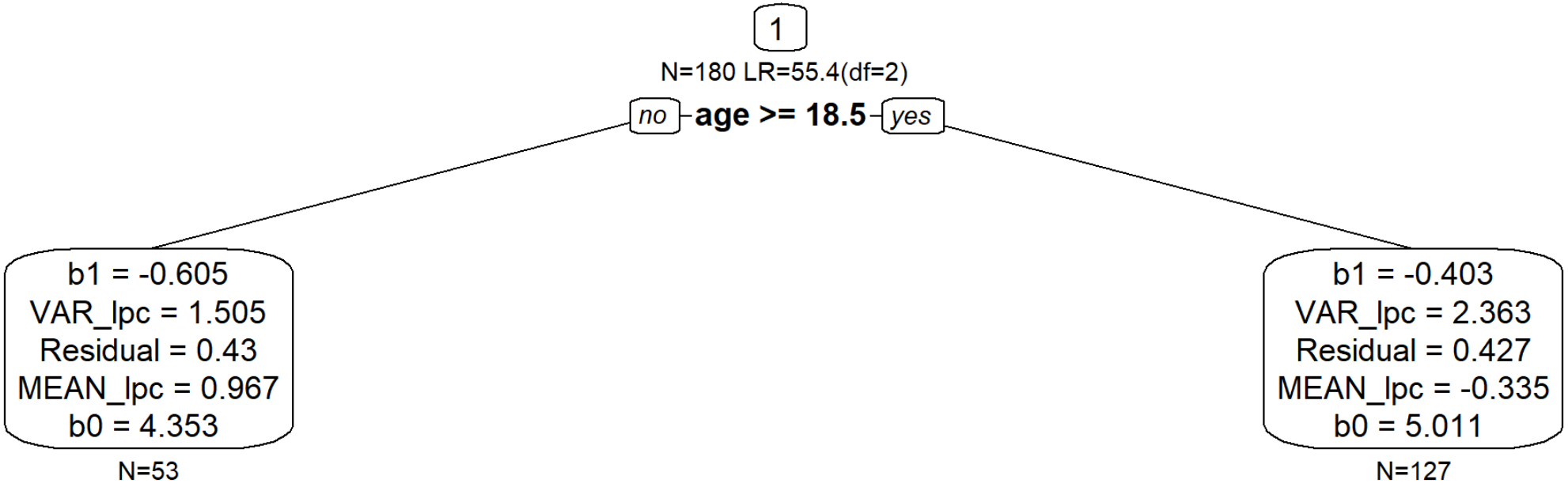
(Subscriptions per price-by-citation in log-scale)

Regression SEM Tree



SEM Tree with focus paramater

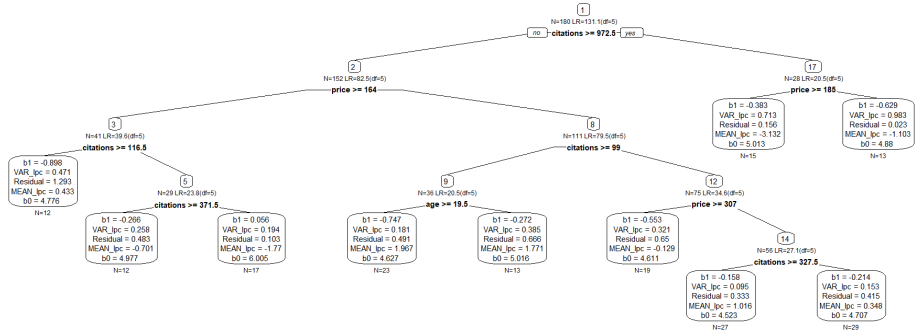
Focus parameter on regression coefficient and constant:



Identical result: somewhat shallower slope for older journals

Tree w/o focus parameter

- If we run the tree without focus parameter, all parameters become potentially relevant
- Splits could be because of differences in intercept, differences in residual variance, or differences in predictor variance



Parameter Estimates of Trees

Here, we find differences w.r.t. all parameters (e.g., predictor mean and variance *MEAN_lpc* and *VAR_lpc*) and also a large range of regression coefficients b_1 ranging from -0.9 to 0.06.

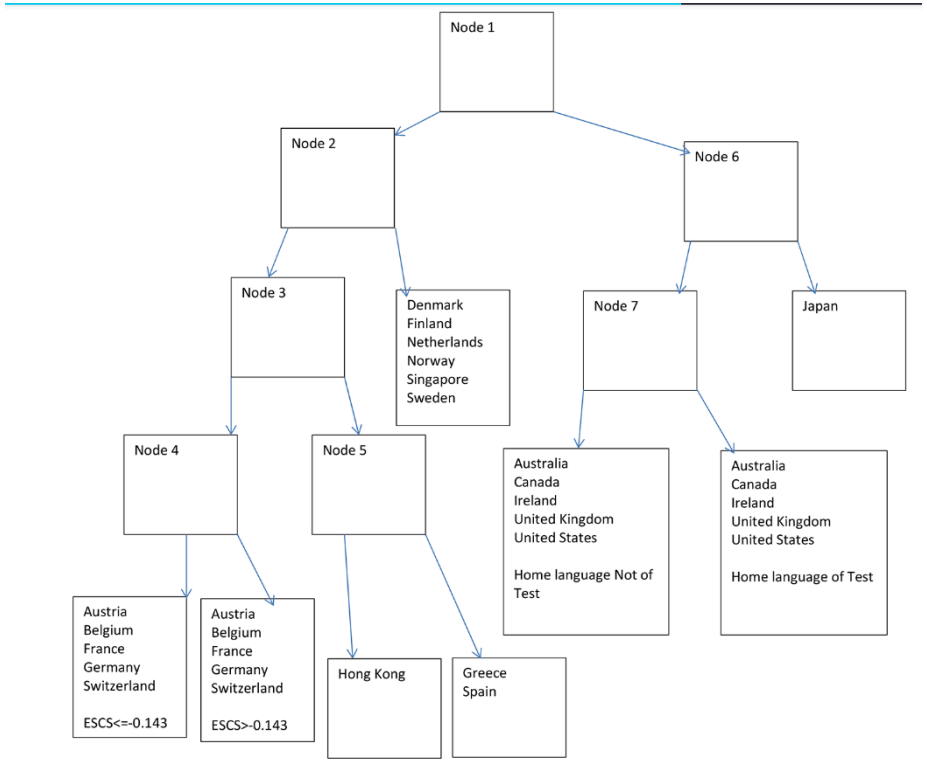
	Leaf #1	Leaf #2	Leaf #3	Leaf #4	Leaf #5	Leaf #6	Leaf #7	Leaf #8	Leaf #9	Leaf #10
b1	-0.90	-0.27	0.06	-0.75	-0.27	-0.55	-0.16	-0.21	-0.38	-0.63
VAR_lpc	0.47	0.26	0.19	0.18	0.38	0.32	0.09	0.15	0.71	0.98
Residual	1.29	0.48	0.10	0.49	0.67	0.65	0.33	0.41	0.16	0.02
MEAN_lpc	0.43	-0.70	-1.77	1.97	1.77	-0.13	1.02	0.35	-3.13	-1.10
b0	4.78	4.98	6.00	4.63	5.02	4.61	4.52	4.71	5.01	4.88

Measurement Invariance

Measurement Invariance Testing

- "Lack of evidence of measurement invariance equivocates conclusions and casts doubt on theory in the behavioral sciences" (Horn and McArdle, 1992)
- Measurement Invariance is usually based on multigrop SEM/CFA (Marsh, Morin, Parker, and Kaur, 2014)
- SEM Trees were suggested as a tool for measurement invariance testing, that is, to explore differential item functioning (Finch, 2017; Sterner and Goretzko, 2023), that is to find moderators of factor loadings
- CFA partitioning can even be used without any groups at all (Merkle and Zeileis, 2013)
- As an example, Finch (2017) looked at a *attitudes towards reading items* questionnaire

Measurement Invariance



SEMtree does not isolate specific differences among model parameters, but rather identifies differences in whole patterns of model parameters - (Finch, 2017)

BFI

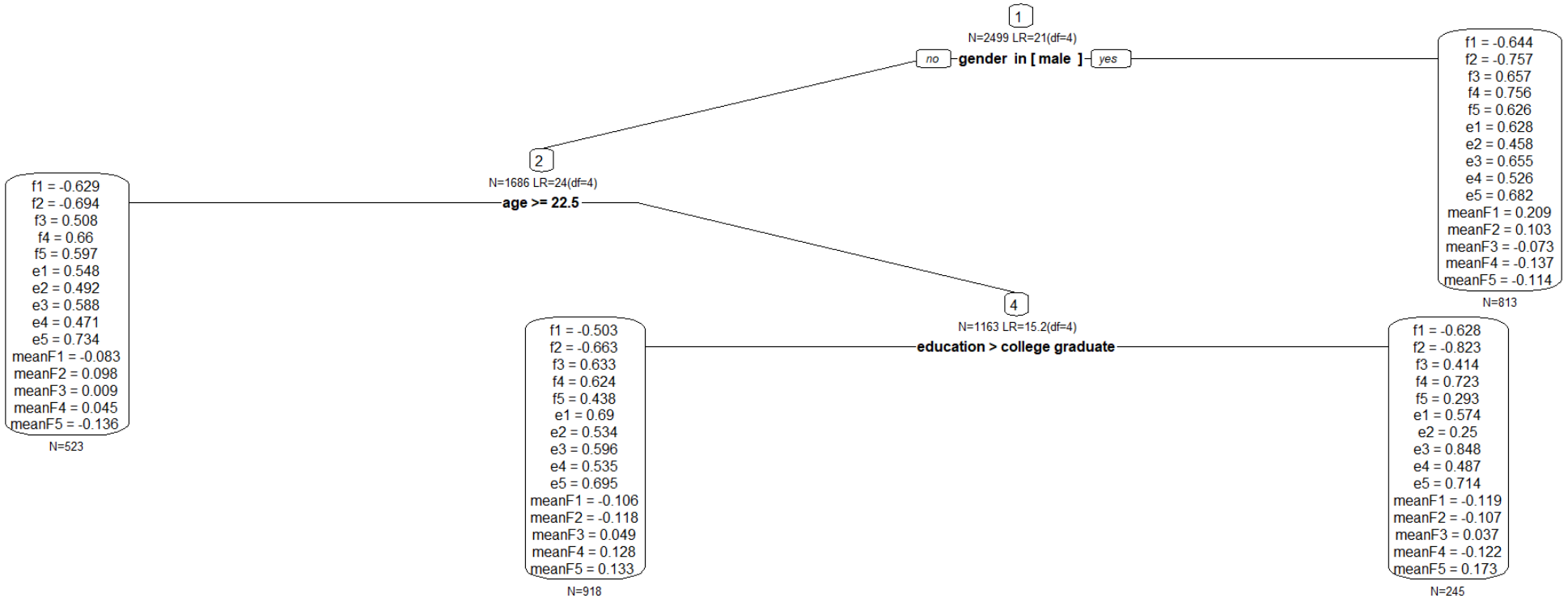
As an example:

- International Personality Item Pool `bfi` from `psych` package
- $n = 2,800$
- 25 personality self report items (representing 5 OCEAN factors)
- Three demographic variables as predictors: age (metric), education (ordinal), sex (nominal)

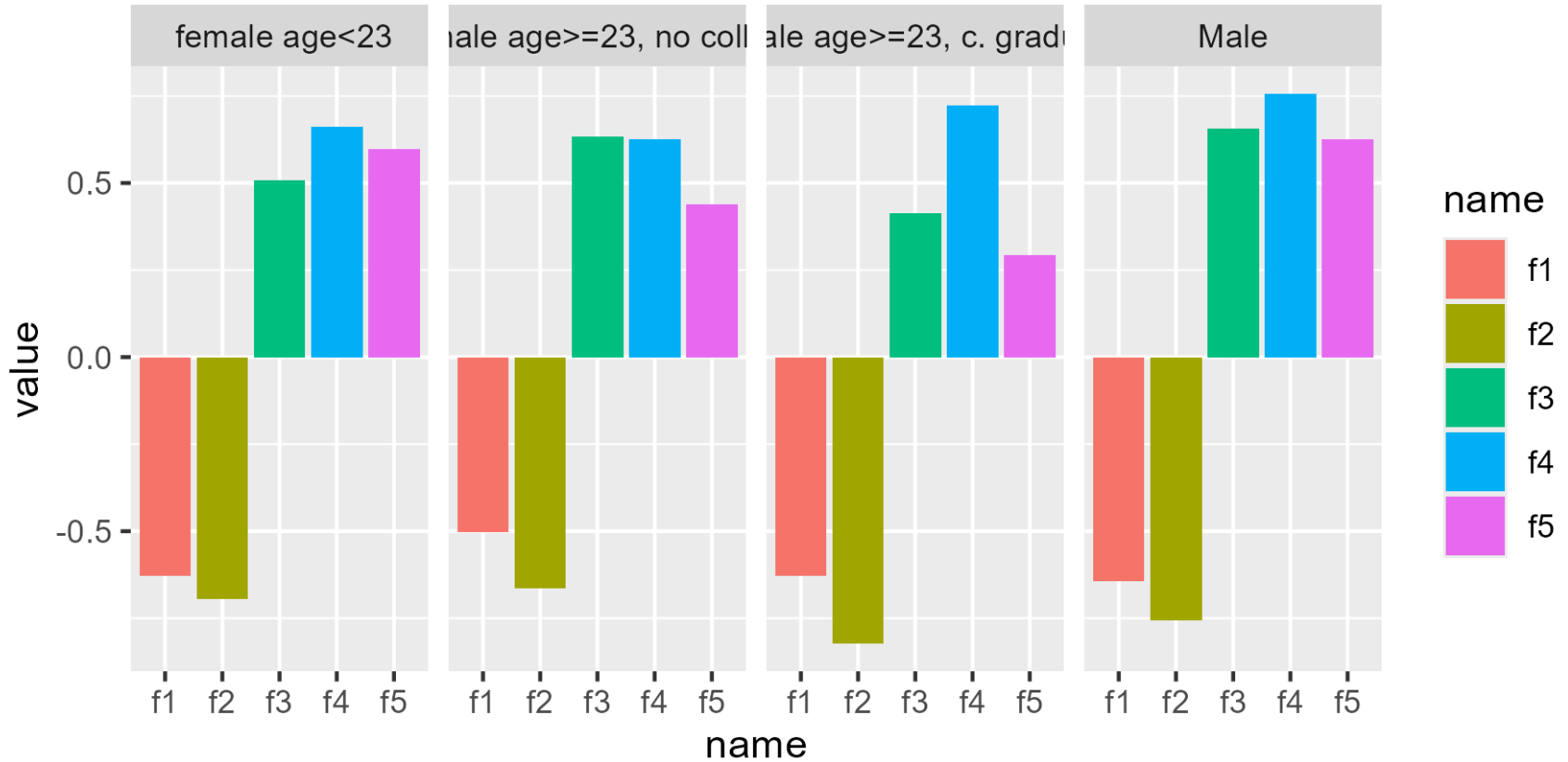
BFI

- Specify a CFA of five indicators for the construct *extraversion*: "don't talk a lot", "find it difficult to approach others", "Know how to captivate propl", "Make friends easily", "Take charge"
- Anchor item: "don't talk a lot"
- choose only factor loadings and residuals as focus parameters (but not intercepts or latent mean/variance)
- Predictors: age, sex, and education

BFI Tree


1
N=2499 LR=21(df=4)
no
gender in [male]
yes
f1 = -0.644
f2 = -0.757
f3 = 0.657
f4 = 0.756
f5 = 0.626
e1 = 0.628
e2 = 0.458
e3 = 0.655
e4 = 0.526
e5 = 0.682
meanF1 = 0.209
meanF2 = 0.103
meanF3 = -0.073
meanF4 = -0.137
meanF5 = -0.114
N=813
2
N=1686 LR=24(df=4)
age >= 22.5
f1 = -0.629
f2 = -0.694
f3 = 0.508
f4 = 0.66
f5 = 0.597
e1 = 0.548
e2 = 0.492
e3 = 0.588
e4 = 0.471
e5 = 0.734
meanF1 = -0.083
meanF2 = 0.098
meanF3 = 0.009
meanF4 = 0.045
meanF5 = -0.136
N=523
4
N=1163 LR=15.2(df=4)
education > college graduate
f1 = -0.503
f2 = -0.663
f3 = 0.633
f4 = 0.624
f5 = 0.438
e1 = 0.69
e2 = 0.534
e3 = 0.596
e4 = 0.535
e5 = 0.695
meanF1 = -0.106
meanF2 = -0.118
meanF3 = 0.049
meanF4 = 0.128
meanF5 = 0.133
N=918
f1 = -0.628
f2 = -0.823
f3 = 0.414
f4 = 0.723
f5 = 0.293
e1 = 0.574
e2 = 0.25
e3 = 0.848
e4 = 0.487
e5 = 0.714
meanF1 = -0.119
meanF2 = -0.107
meanF3 = 0.037
meanF4 = -0.122
meanF5 = 0.173
N=245

BFI Loadings

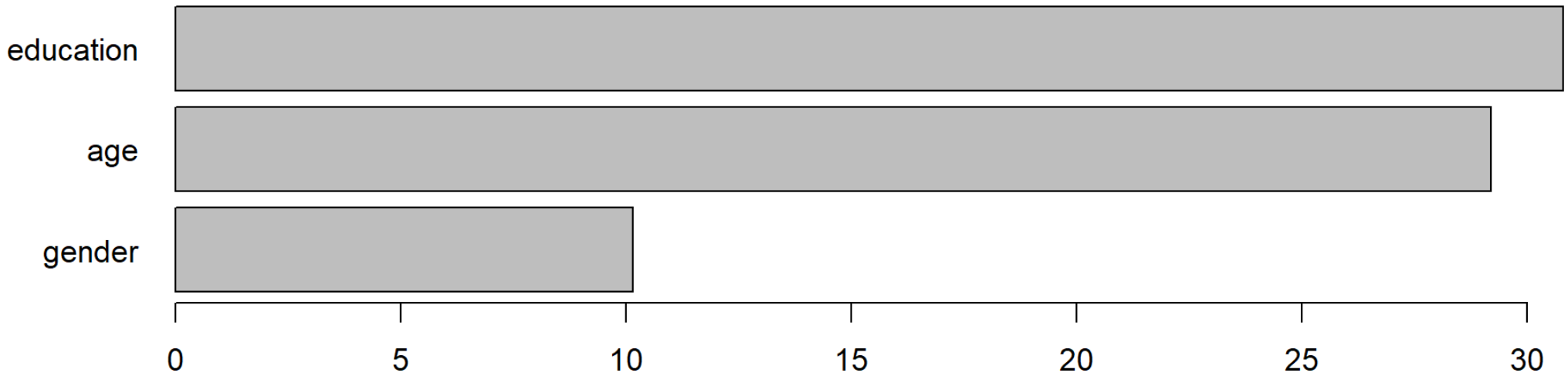


(f3: Know how to captivate people and f5: Take charge - professional skills vs personality?)

BFI Forest

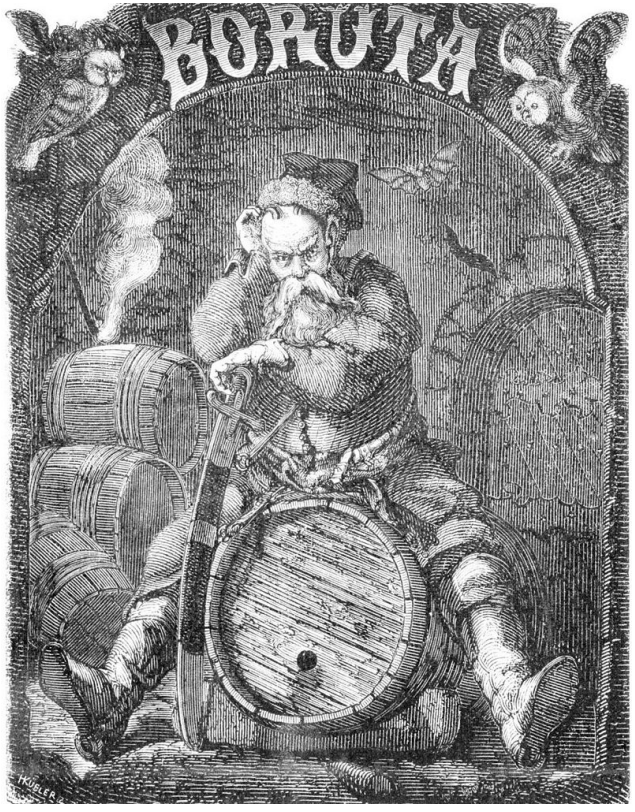
Permutation variable importance with focus parameters loadings:

```
vim <- varimp(frst, method="permutationFocus")
```



Outlook I

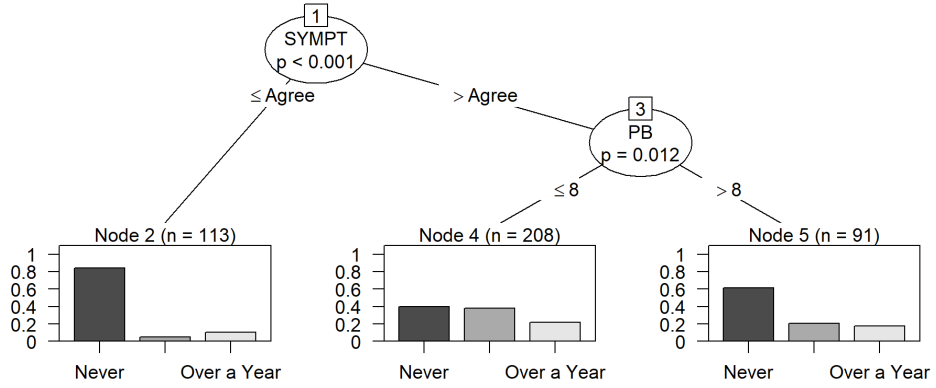
- Adapte BORUTA (Kursa and Rudnicki, 2010a) to SEM Tree (with Priyanka Paul & Timothy Brick, PennState)
- BORUTA is a heuristic to determine a cut-off between important and not-important variables



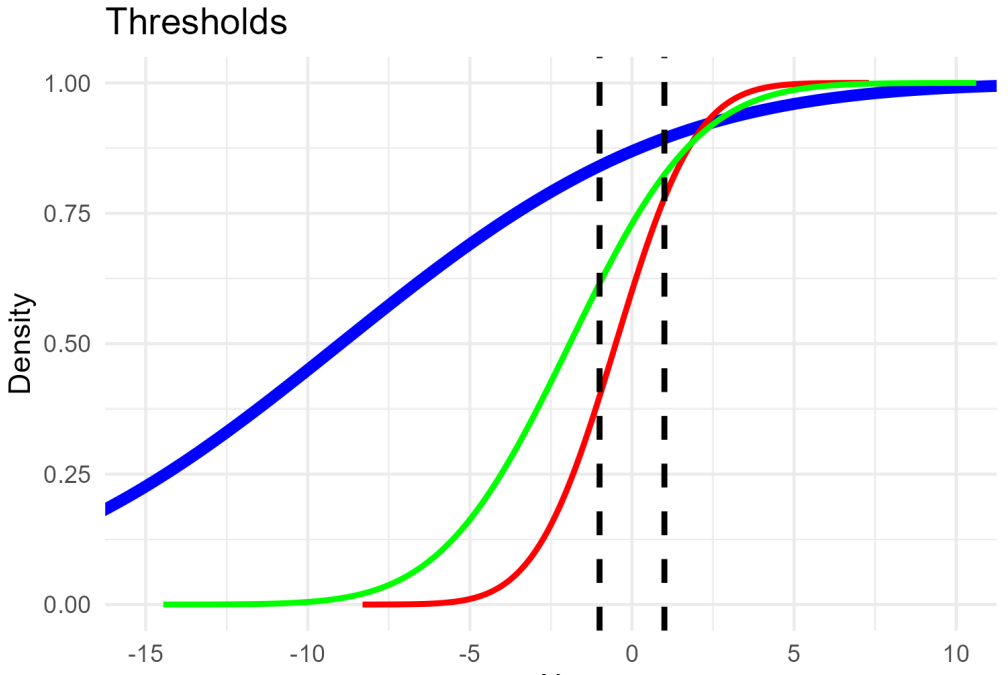
A spirit or devil from slavic mythology, image from Wikipedia/Public

Outlook II

(Joint) Ordinal models using threshold models (example: mammography screening experience and opinions from (Hothorn, Hornik, and Zeileis, 2006) with $n = 412$)



- SYMPT: You do not need a mamogram unless you develop symptoms
- PB: Perceived benefit (low value = strong



Summary

- SEM Trees and Forests are a form of model-based recursive partitioning
- They can be implemented either via the `semtree` package or via `partykit` (https://www.zeileis.org/news/lavaantree/)
- Offer a way to explore theory-driven SEM with variable importance and partial dependence plots
- Focus parameters may be important when exploring the importance of predictors for subsets of parameters (e.g., individual differences vs means, or intercept vs slopes in growth models)
- If you have a theory, test the theory first, then explore!

Thank You

- Slides: https://github.com/brandmaier/focus-talk-dagstat2025
- Package on CRAN: `semtree`

Contact:
andreas.brandmaier@medicalschool-berlin.de or @brandmaier.bsky.social on Bluesky or
https://www.brandmaier.de

Reproducibility + Exploratory Methods

Why should we work reproducibly?

Many good reasons like:

- Transparency
- Trustworthiness
- Replication
- Cumulative science

> Transparency and accessibility are central scientific values, and open, reproducible projects will increase the efficiency and veracity of knowledge accumulation .

Your closest collaborator is you six months ago, but you don’t reply to emails.

From Karl Broman's lecture on reproducibility, paraphrasing Mark Holden

Forensics

If an analysis is repeated later in time and results do not align with what was published, it could be due to:

- changes in the core functions of dependent packages (e.g., bugfixes)
- changes in defaults of dependent packages (e.g., default test statistic for splits, default stopping criteria, default bucket size)
- changes in the data used (e.g., preprocessing, outlier removal)
- changes in the R script used (e.g., multiple versions were created during development and it's unclear, which one was used ultimately)

Sources of Failure to Reproduce Results

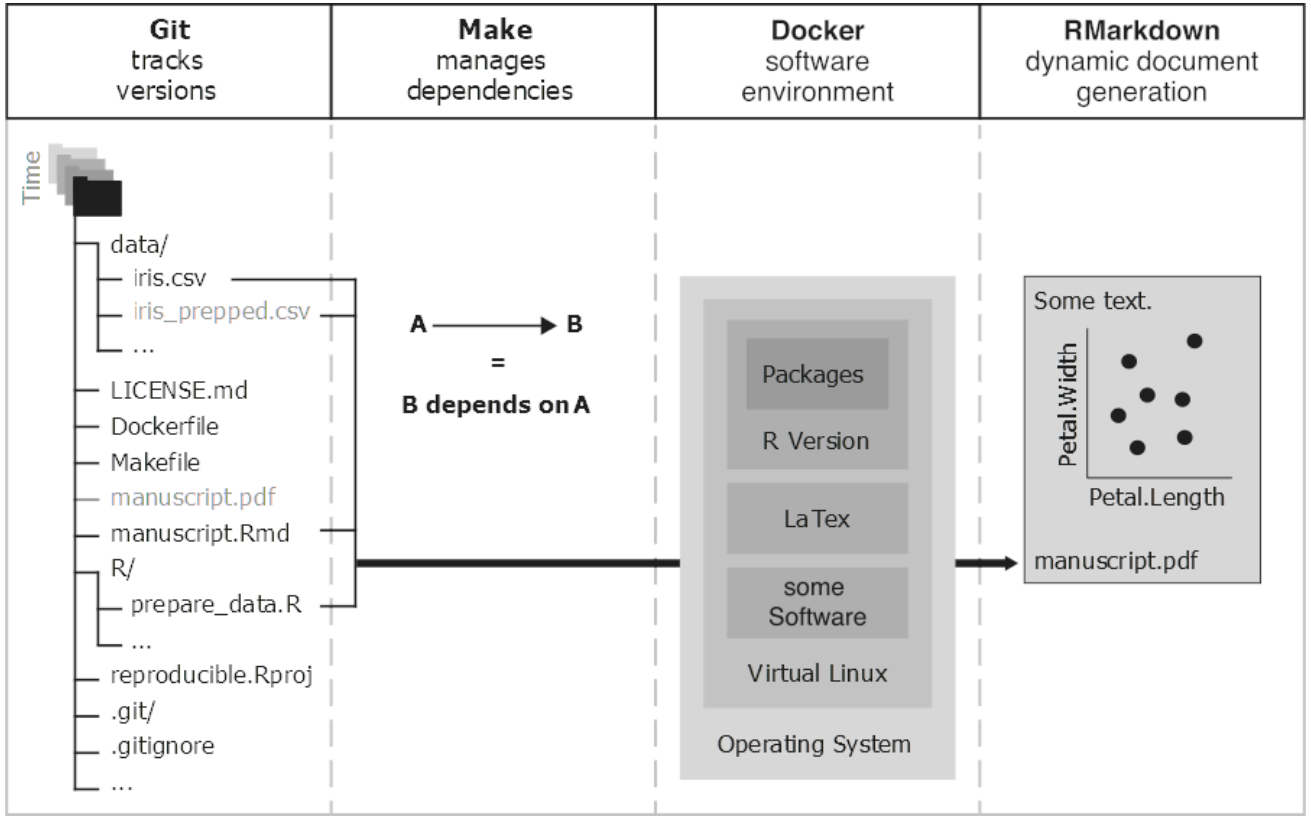
1. **Multiple versions of scripts/data** (e.g., dataset has changed over, i.e., was further cleaned or extended)

2. **Multiple scripts** in a pipeline; unclear which scripts should be executed in which order

3. **Copy&paste errors** (e.g., inconsistency between reported result and reproduced result)

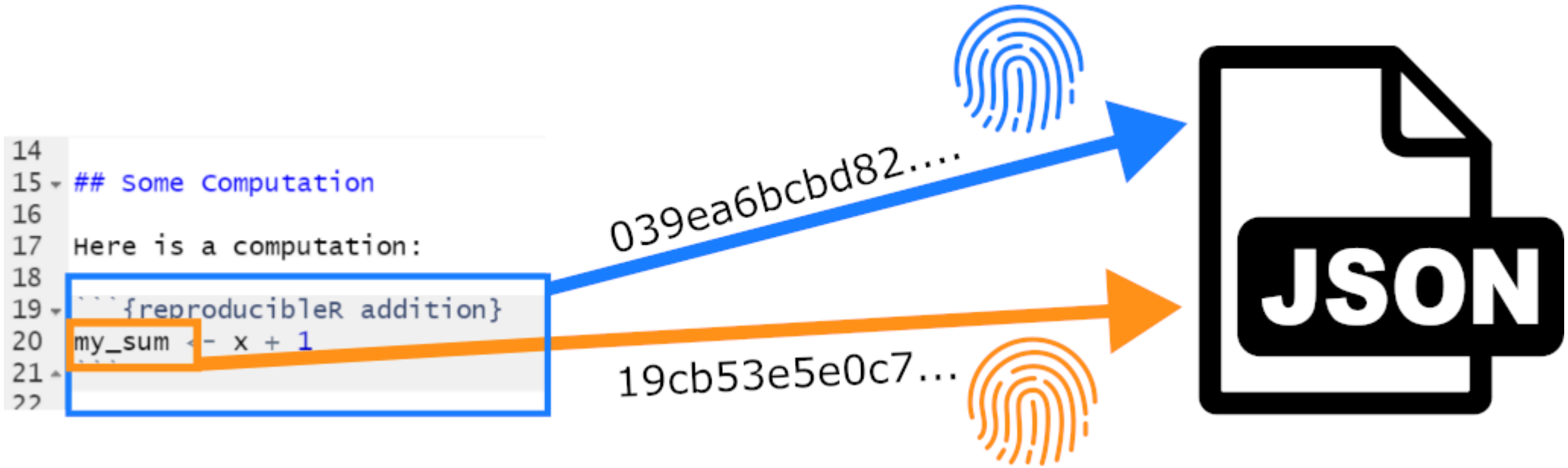
4. Broken **software dependencies** (e.g., analysis broken after update, missing package, just comes out differently on a different computer)

Four Elements of Reproducibility



from Peikert and Brandmaier (2020)

reproducibleRchunks



(now on CRAN: reproducibleRchunks)

References

Arnold, M., M. C. Voelkle, and A. M. Brandmaier (2021). "Score-guided structural equation model trees". In: *Frontiers in psychology* 11, p. 564403.

Brandmaier, A. M., T. von Oertzen, J. J. McArdle, et al. (2013). "Structural equation model trees." In: *Psychological methods* 18.1, pp. 71-86.

Brandmaier, A. M., J. J. Prindle, J. J. McArdle, et al. (2016). "Theory-guided exploration with structural equation model forests". In: *Psychological Methods* 21.4, pp. 66–-582.

Brandmaier, A. M., N. Ram, G. G. Wagner, et al. (2017). "Terminal decline in well-being: The role of multi-indicator constellations of physical health and psychosocial correlates." In: *Developmental Psychology* 53.5, pp. 996-1012.

Finch, W. H. (2017). "Structural equation modelling trees for invariance assessment". In: *International Journal of Quantitative Research in Education* 4.1-2, pp. 72-93.

Gigerenzer, G. and S. Kurzenhaeuser (2005). "Fast and frugal heuristics in medical decision making".

Horn, J. L. and J. J. McArdle (1992). "A practical and theoretical guide to measurement invariance in aging research". In: *Experimental aging research* 18.3, pp. 117-144.

Hothorn, T., K. Hornik, and A. Zeileis (2006). "Unbiased recursive partitioning: A conditional inference framework". In: *Journal of Computational and Graphical statistics* 15.3, pp. 651-674.

Kursa, M. B. and W. R. Rudnicki (2010a). "Feature Selection with the Boruta Package". In: *Journal of Statistical Software* 36.11, pp. 1-13. URL: https://doi.org/10.18637/jss.v036.i11.

Marsh, H. W., A. J. Morin, P. D. Parker, et al. (2014). "Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis". In: *Annual review of clinical psychology* 10.1, pp. 85-110.

Merkle, E. C. and A. Zeileis (2013). "Tests of measurement invariance without subgroups: A generalization of classical methods". In: *Psychometrika* 78.1, pp. 59-82.

Sterner, P. and D. Goretzko (2023). "Exploratory factor analysis trees: Evaluating measurement invariance between multiple covariates". In: *Structural Equation Modeling: A Multidisciplinary Journal* 30.6, pp. 871-886.

Strobl, C., A. Boulesteix, T. Kneib, et al. (2008). "Conditional variable importance for random forests". In: *BMC bioinformatics* 9, pp. 1-11.

Strobl, C., A. Boulesteix, A. Zeileis, et al. (2007). "Bias in random forest variable importance measures: Illustrations, sources and a solution". In: *BMC bioinformatics* 8, pp. 1-21.

Zeileis, A., T. Hothorn, and K. Hornik (2008). "Model-based recursive partitioning". In: *Journal of Computational and Graphical Statistics* 17.2, pp. 492-514.