KAMUZU UNIVERSITY OF HEALTH SCIENCES

Selection of Variables and Functional Forms for Multivariable Models.

Marc Henrion

For the Topic Group 2 of the STRATOS Initiative

Slides adapted from material provided by Georg Heinze, Willi Sauerbrei and from discussions within the STATOS TG2 group.

IBS SUSAN 2023 – Kumasi, Ghana
10 September 2023

The STRATOS initiative
STRengthening Analytical Thinking for Observational Studies

STRATOS INITIATIVE

LSTM

wellcome

Preliminaries

GitHub repo for this workshop:

https://github.com/gitMarcH/IBS_Ghana_StratosTG2

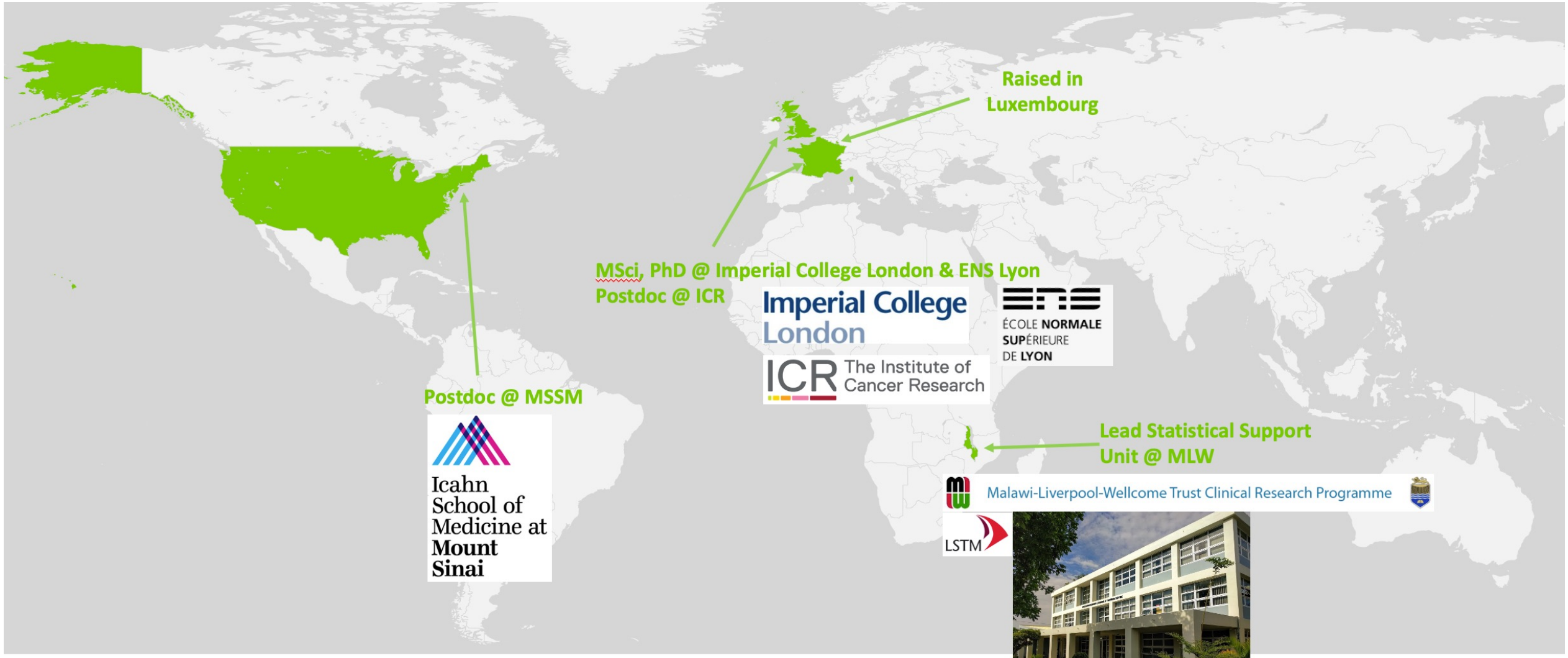
For the examples:
This workshop uses R and you will need the following packages:

- tidyverse (optional)
- MASS
- glmnet
- splines
- rms
- mfp

Install them now!

Introduction

Introduction: About myself



The STRATOS Initiative

https://www.stratos-initiative.org

STRATOS INITIATIVE

STRengthening Analytical Thinking for Observational Studies

Accessible and Accurate Guidance in the Design and Analysis of Observational Studies

Context

- Observational studies pose many design and statistical challenges
- Valid observational research depends on careful study design, high data quality, appropriate statistical methods and accurate interpretation of results

The Problem

- Statistical methods has seen exponential advancements
  - diffusion of methodological innovation is slow
  - many developments are not applied in practice
- Even worse, 'standard' analyses reported in the medical literature are often based on unrealistic assumptions or use inappropriate methods, casting doubt on their results and conclusions
- Analysts, reviewers, editors, readers and many more stakeholders and consumers need guidance for key issues in the design and analysis of observational studies

Filling the Gap: The STRATOS initiative

- The STRenghtening Analytical Thinking for Observational Studies (STRATOS) initiative was launched in August 2013

  In November 2019
  - ~100 researchers from 18 countries worldwide with background in biostatistical and epidemiological methods
  - Connected with leading international organisations, e.g.,
    - International Society of Clinical Biostatistics (ISCB), and
    - International Biometric Society (IBS)
- Website includes member details, publications & resources http://www.stratos-initiative.org/

STRATOS Objectives

- Provide accessible and evidence-based guidance for key topics in the design and analysis of observational studies
- Guidance is intended for applied statisticians and other data analysts with varying levels of statistical education, experience and interests

Organisational Structure

STRATOS is led by an **Executive Committee** and a **Steering Group**. It has the following topic groups and cross-cutting panels:

Topic Groups (TGs)	
1	Missing data
2	Selection of variables and functional forms in multivariable analysis
3	Initial data analysis
4	Measurement error and misclassification
5	Study design
6	Evaluating diagnostic tests and prediction models
7	Causal Inference
8	Survival analysis
9	High-dimensional data

Panels	
MP	Membership
PP	Publications
GP	Glossary
WP	Website
RP	Literature Review
BP	Bibliography
SP	Simulation Studies
DP	Data Sets
TP	Knowledge Translation
CP	Contact Organizations
VP	Visualisation

Membership:

Regular: Experienced researchers who contribute regularly to STRATOS TG +/- panel activities

Experienced Adjunct: Experienced researchers who contribute occasionally to STRATOS activities

Early Career Adjunct: Researchers at early career stages who participate in TG +/- panel activities

Clinical Affiliates: Clinicians engaged in research with interest and practical experience with statistics

Framework for Guidance Development

Guidance is aimed at users with three levels of statistical knowledge:

Level 1: Low statistical knowledge
- Propose acceptable methods that are easily implemented
- Highlight weaknesses of common approaches

Level 2: Experienced statistician
- Refer to advantages and disadvantages of competing approaches
- Propose advanced methodology feasible by experienced analysts

Level 3: Expert in a specific area
- Consider recent developments with statements about possible advantages and disadvantages
- Identify areas needing more methodological research or guidance

Stages in Guidance Development

Phase I: Experts need to work on state-of-the-art methods and develop guidance for knowledge level 2. Comparisons (simulations) required!

Phase II: Extend guidance to level 1 (simpler methods but still acceptable) and experts work on improved methodology aiming to improve level 2 guidance

Activities & Selected Publications

Important Meetings/Mini-Symposia

- ISCB conferences (2013, 2014, 2015, 2016, 2018, 2019, 2020)
- IBS - Invited sessions at IBC (2016, 2020)
  - Meetings of regional groups of IBS (2017, 2018, 2020)
- Banff International Research Station, Canada (general meetings in 2016, 2019)

Selected Publications

- Sauerbrei *et al* for the STRATOS initiative. STRengthening Analytical Thinking for Observational Studies: the STRATOS initiative. *Stat Med* 2014; 33:5413-5432
- Huebner *et al* for TG3. A contemporary conceptual framework for initial data analysis. *Obs Stud* 2018; 4: 171-192
- Shaw *et al* for TG4. Epidemiologic analyses with error-prone exposures: review of current practice and recommendations. *Ann Epidemiol* 2018; 28: 821–828
- Boulesteix *et al* for the Simulation Studies panel. On the necessity and design of studies comparing statistical methods. *Biom J* 2018; 60: 216-218
- **Series of short papers** from the TGs in the Biometric Bulletin of IBS (available on the STRATOS website) provide an overview of TG activities.

Statistical methodology – current situation

- Substantial development of statistical methodology in past 20+ years.
- Largely driven by:
  1. Computational feasibility
     - Complex, sophisticated analyses possible that were not feasible some years ago.
     - Resampling and Bayesian methods key examples.
  2. Available software
     - Open source software platforms such as R, python and others allows rapid implementation and adoption of new statistical methodology.

Unfortunately, many sensible improvements are ignored in practical statistical analyses

Statistical methodology – current situation

- Overwhelming concern **with theoretical aspects**.

- Very **limited guidance** on key issues that are vital **in practice**, discourages analysts from utilizing more sophisticated and possibly more appropriate methods in their analyses

Statistical methodology – problem well known

The Economist ‘*Unreliable research: Trouble at the lab.*’ (October 2013):

“Scientists’ grasp of statistics has not kept pace with the development of complex mathematical techniques for crunching data. Some scientists use inappropriate techniques because those are the ones they feel comfortable with; others latch on to new ones without understanding their subtleties. Some just rely on the methods built into their software, even if they don’t understand them.”

The Economist | Menu | Weekly edition | The world in brief | Search

Briefing | Unreliable research

Trouble at the lab

Scientists like to think of science as self-correcting. To an alarming degree, it is not

STRATOS – improvements needed

At least two tasks are essential:

1. **Experts** in specific methodological areas have to work towards **developing guidance.**

2. An ever-increasing need for **continuing education** at all stages of the career:
   - For busy applied researchers it is often difficult to follow methodological progress even in their principal application area
     - Reasons are diverse.
     - Consequence is that analyses are often deficient.
   - **Knowledge** gained through research on statistical methodology needs to be **transferred** to the broader community.
   - Many **analysts** would be **grateful for** an overview on the current **state of the art** and for **practical guidance**.

STRATOS – aims of the initative

- **Provide evidence supported guidance** for highly relevant issues in the design and analysis of observational studies.
- As the **statistical knowledge** of the analyst **varies** substantially, guidance has to keep this background in mind. **Guidance** has to be provided **at several levels**.
- For the **start** we will concentrate on **state-of-the-art** guidance and the necessary evidence.
- Help to identify questions requiring much more primary research

The overarching long-term aim is to improve key parts of design and statistical analyses of observational studies in practice.

STRATOS – different levels of knowledge

Level 1: Low statistical knowledge

- Most analyses are done by analysts at that level.
- Point out weaknesses of approaches often used despite of problems (e.g. categorizing continuous variables in the analysis; complete case analysis if a case has missing values in one or more variables, …).
- Propose methods which may not be optimal or state-of-the-art, but which are easy to use and which are still acceptable from a methodological point of view.
- Software should be generally available.

STRATOS – different levels of knowledge

Level 2: Experienced statistician

- Methodology perhaps slightly below state of the art, but doable by every experienced analyst.
- Advantages and disadvantages of competing approaches, point to the importance and implications of underlying assumptions.
- Sufficient guidance about software plays a key role that the approaches are also used in practice.

STRATOS – different levels of knowledge

Level 3: Expert in a specific area

Experts aim to improve statistical models and to adapt them to complex real problems, or to develop new and more complicated approaches.

Issues:

- Advantages and usefulness in practice are not always clear.
- Illustration on a small number of examples in specific situations.
- Software requires specific knowledge and may not be generally available (less and less the case!).

Opportunities:

- Overview of recent research with statements about possible advantages and disadvantages is needed.
- Could help to identify important weaknesses of Level 2 proposals.
- Help to identify areas needing more methodological research.
- Trigger the development of software for more general use.

STRATOS – topic groups & panels

	Topic Groups (TGs)
1	Missing data
2	Selection of variables and functional forms in multivariable analysis
3	Initial data analysis
4	Measurement error and misclassification
5	Study design
6	Evaluating diagnostic tests and prediction models
7	Causal Inference
8	Survival analysis
9	High-dimensional data

	Panels
MP	Membership
PP	Publications
GP	Glossary
WP	Website
RP	Literature Review
BP	Bibliography
SP	Simulation Studies
DP	Data Sets
TP	Knowledge Translation
CP	Contact Organizations
VP	Visualisation

STRATOS Topic Group 2

STRATOS TG2

Selection of variables and functional forms in multivariable analysis

Chairs

- Georg Heinze
- Aris Perepeglou
- Willi Sauerbrei

Focus

- Identify influential variables.
- Gain insight into their individual & joint relationship with the outcome.

Interrelated challenges (Harrell 2001, Sauerbrei et al. 2007)

- Selection of variables for inclusion in a multivariable explanatory model.
- Choice of the functional forms for continuous variables.

STRATOS TG2 – first challenge

Multivariable models typically built through a combination of

- A priori inclusion of well established 'predictors' of the outcome of interest, and
- A posteriori selection of variables (often using arbitrary, data-dependent procedures)

Examples

- Univariable association tests.
- Forward variable selection.
- Backward elimination.

Consensus that all of the many suggested model building strategies have weaknesses (Miller 2002) but opinions on the relative advantages and disadvantages of particular strategies differ considerably.

More advanced methods exist (e.g. regularization techniques, resampling based methods, …), BUT

- No clear optimal method → need neutral, systematic comparisons
- Need for clearer guidance.

→ **STRATOS TG2**

STRATOS TG2 – second challenge

The effects of continuous predictors are typically modeled by

- Categorizing.
- Assuming linear relationships with the outcome (possibly after simple transformations).

Problematic if:

- Reasons for choosing such conventional representation of continuous variables are not discussed
- Validity of the underlying assumptions is not assessed.

Flexible modeling techniques have been developed:

- Fractional polynomials (Royston and Altman 1994, Royston and Sauerbrei 2008)
- Splines (many 'flavours'; Boer 2001, Harrell 2001, Wood 2006, Hastie and Tibshirani 1990)

For multivariable analysis, these smoothers have been incorporated in generalized additive models.

Unclear:

- Optimal approach → need neutral, systematic comparisons.
- Need for clearer guidance.

→ STRATOS TG2

STRATOS TG2

Further practical guidance is urgently needed:

- Extended investigations of analytical properties
- Systematic comparisons between alternative methods.

TG2 aims to:

1. Do comprehensive (and ongoing) reviews to
   - Identify and assess methods currently used in practice.
   - Find any published guidelines on selection of variables and their functional forms.
   - Find systematic simulation-based comparisons of alternative techniques, especially in multivariable analyses.

   (Where lacking, TG2 initiates such comparison projects – e.g. current COVID-19 modelling review project with focus on model building)
2. Develop consensus-based recommendations and systematic guidance to address accuracy, efficiency, transportability, ease of implementation and interpretability, in wide range of applications (Sauerbrei 2007).

3 levels of guidance: Level 1 (low stats level), Level 2 (experienced statisticians), Level 3 (experts in area)

STRATOS TG2

- **We assume ‘homework’ has been done:**
  - Initial data analysis (TG3)
    - Missing values, univariate X, multivariate X, keep outcome Y separate!
  - Reasonable model class was chosen
    - Linear, binary, multinomial, censoring, …

- **Which variables to include?**
  Driven by the interpretation of the model as:
  - Description
  - Prediction
  - Causal explanation

- **How to specify how to model continuous variables (the ‚functional form‘)?**

Variable selection

The aims of modelling

The aims of modelling

- Describe?

The aims of modelling

- Describe?

- Explain?

The aims of modelling

- Describe?
- Explain?
- Predict?

The aims of modelling

- Describe?
- Explain?
- Predict?

A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks

Significance, 2019

Miguel A. Hernán, John Hsu, and Brian Healy

Table 1—Examples of Tasks Conducted by Data Scientists Working with Electronic Health Records

	Data Science Task		
	Description	**Prediction**	**Causal inference**
Example of scientific question	How can women aged 60–80 years with stroke history be partitioned in classes defined by their characteristics?	What is the probability of having a stroke next year for women with certain characteristics?	Will starting a statin reduce, on average, the risk of stroke in women with certain characteristics?
Data	• Eligibility criteria • Features (symptoms, clinical parameters …)	• Eligibility criteria • Output (diagnosis of stroke over the next year) • Inputs (age, blood pressure, history of stroke, diabetes at baseline)	• Eligibility criteria • Outcome (diagnosis of stroke over the next year) • Treatment (initiation of statins at baseline) • Confounders • Effect modifiers (optional)
Examples of analytics	Cluster analysis …	Regression Decision trees Random forests Support vector machines Neural networks …	Regression Matching Inverse probability weighting G-formula G-estimation Instrumental variable estimation …

Statistical Science
2010, Vol. 25, No. 3, 289–310
DOI: 10.1214/10-STS330
© Institute of Mathematical Statistics, 2010

To Explain or to Predict?

Galit Shmueli

I define *predictive modeling* as the process of applying a statistical model or data mining algorithm to data for the purpose of predicting new or future observations. In particular, I focus on nonstochastic prediction

a causal theory is much weaker. Hence, given this reality, I define *explaining* as causal explanation and *explanatory modeling* as the use of statistical models for testing causal explanations.

Although not the focus of this article, a third type of modeling, which is the most commonly used and developed by statisticians, is descriptive modeling. This type of modeling is aimed at summarizing or representing the data structure in a compact manner. Unlike explanatory modeling, in descriptive modeling the reliance on an underlying causal theory is absent or incorporated in a less formal way. Also, the focus is at the

To describe, predict or explain?

- **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

- **Prediction models**
  - Interest in accurate predictions for future application
  - No concern about causality and confounding (association)
  - black-box vs. transparent
  - Prognostic and diagnostic prediction models

- **Explanatory models**
  - Interest in effect of an intervention on patient outcomes
  - Often achieved by counterfactual prediction
  - Confounder selection

Model building & modelling aims

- **Description**:
  - Just X and Y: understand how Y is associated with X‘s
  - Simple: make general, widely valid statements about these associations
  - Often misspecified “by intention”

- **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 and poor calibration

- **Explanation** (causal inference):
  - Main concern: correct adjustment for confounders
  - Misspecification leads to biased effect estimate
  - Simplicity not ultimately needed; may reduce variance

Model building & modelling aims

Sauerbrei *et al. Diagnostic and Prognostic Research* (2020) 4:3
https://doi.org/10.1186/s41512-020-00074-3

Diagnostic and Prognostic Research

COMMENTARY Open Access

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

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

Variable selection

- A lot of methods.
- A lot of opinions.

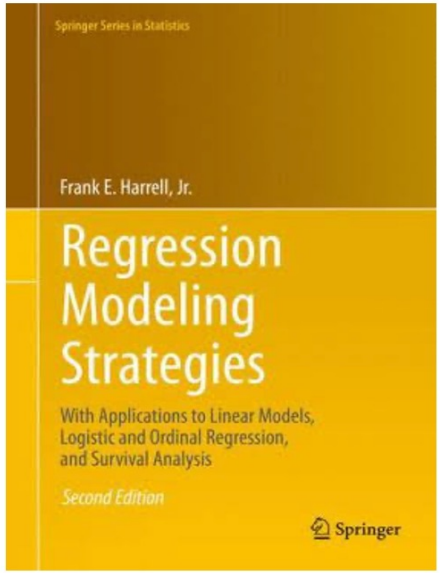
Variable selection

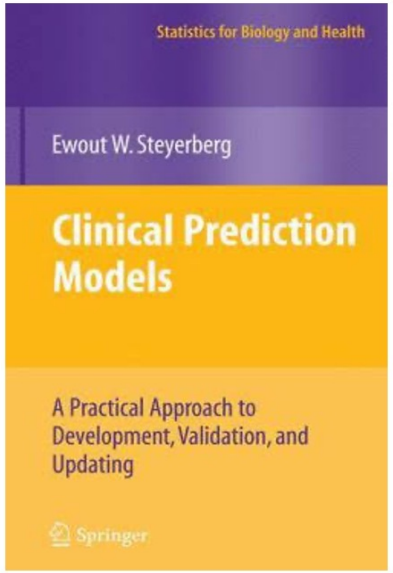
Harrell (2001)

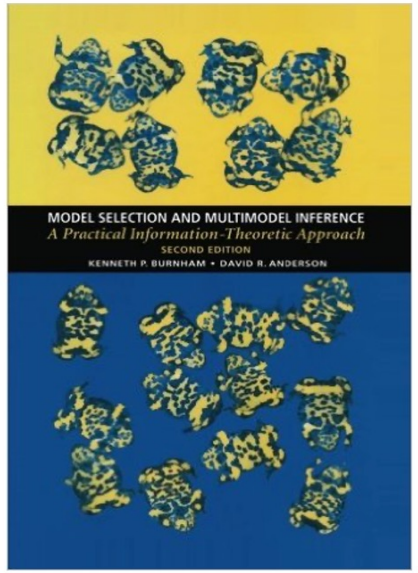
Steyerberg (2009)

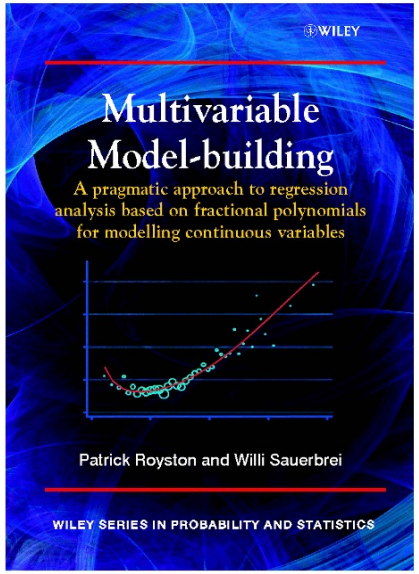
Burnham & Anderson (2002)

Royston & Saueberei (2008)









Traditional & modern variable sel. methods

• Univariate selection	$\alpha = 0.05, 0.1, 0.2, \ldots$
• Forward selection	$\alpha = 0.05, 0.1, AIC, 0.2, \ldots$
• Backward elimination	$\alpha = 0.05, 0.1, AIC$
• All subsets methods → criteria dependent	Mallows Cp, AIC, BIC, …
• Change-of-estimate based	$\Delta\hat{\beta} = 5\%,\ \Delta\hat{\beta} = 10\%$
• Augmented backward elimination*	$\alpha = 0.157, \tau = 0.05$
• Lasso	λ selected by AIC, λ cross-validated

*Dunkler et al, 2014

In practice often a combination, e.g. univariate + backward:

"From the variables that were associated with Y in univariate models (Table 2), XX and XY were kept as independent predictors in the model…"

Stepwise procedures

Central Issue: significance level choice depends on aim of the study

Criticism

- FS start with ‚bad‘ univariate models (underfitting).
- BE starts with the full model (overfitting).
- Less critical.
- Multiple testing, p-values incorrect .
- Nevertheless very popular in practice.

State of the art?

We do not know any recommendation which is supported by good evidence from theory or meaningful simulation studies

(Data-driven) variable selection vs. pre-specification

- **Descriptive models**
  - Pre-specify – if we want to describe the data in that way
  - Variable selection – to identify the main associations ('remove noise')?

- **Prediction models**
  - Pre-specify – predictors chosen based on availability, costs, accuracy, reliability, …
  - Variable selection – to decrease prediction error by removing noisy inputs?

- **Explanatory models**
  - Pre-specify – select confounders based on strong assumptions (positivity, DAGs, …)
  - Variable selection – to decrease MSE of estimator?

(Data-driven) variable selection vs. pre-specification

- **Descriptive models**
  - Pre-specify – if we want to describe the data in that way
  - Variable selection – to identify the main associations ('remove noise')?

- **Prediction models**
  - Pre-specify – predictors chosen based on availability, costs, accuracy, reliability, …
  - Variable selection – to decrease prediction error by removing noisy inputs?

- **Explanatory models**
  - Pre-specify – select confounders based on strong assumptions (positivity, DAGs, …)
  - Variable selection – to decrease MSE of estimator?

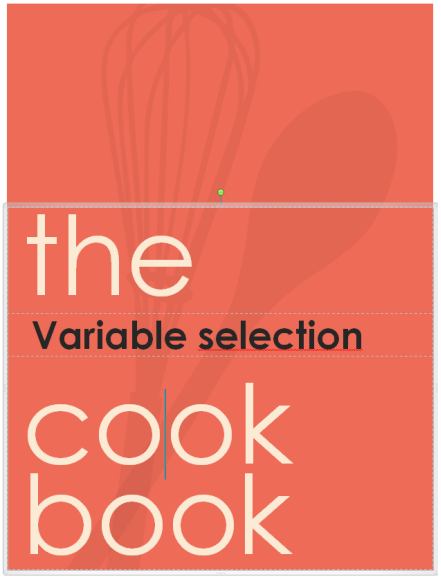
BIAS & VARIANCE

Dos and don’ts

- Don’ts:
  - Perform excessive data exploration (check associations of X‘s with Y)
    - Variable selection algorithms are a cascade of such exploration steps!
  - Poor reporting of what you‘ve done
  - Report final model as if it was prespecified (low standard errors)
  - Misinterpret results
    (‚chest is a predictor of body fat,
    thigh is no predictor of body fat‘)

- Dos:
  - (Will get there in a few slides)

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.

Model stability & bootstrap

- Whichever variable selection method we use: variables either get or do not get selected when we build a model.

- If we'd repeat the experiment and get another dataset, we most likely will end up with a slightly different set of variables.

- We can mimic this by resampling the data we have using the bootstrap.
  (this is just one way of doing this!)

- We can use this to report how stable the variable selection this: which variables get consistently included and which ones only occasionally.

Consequences of variable selection

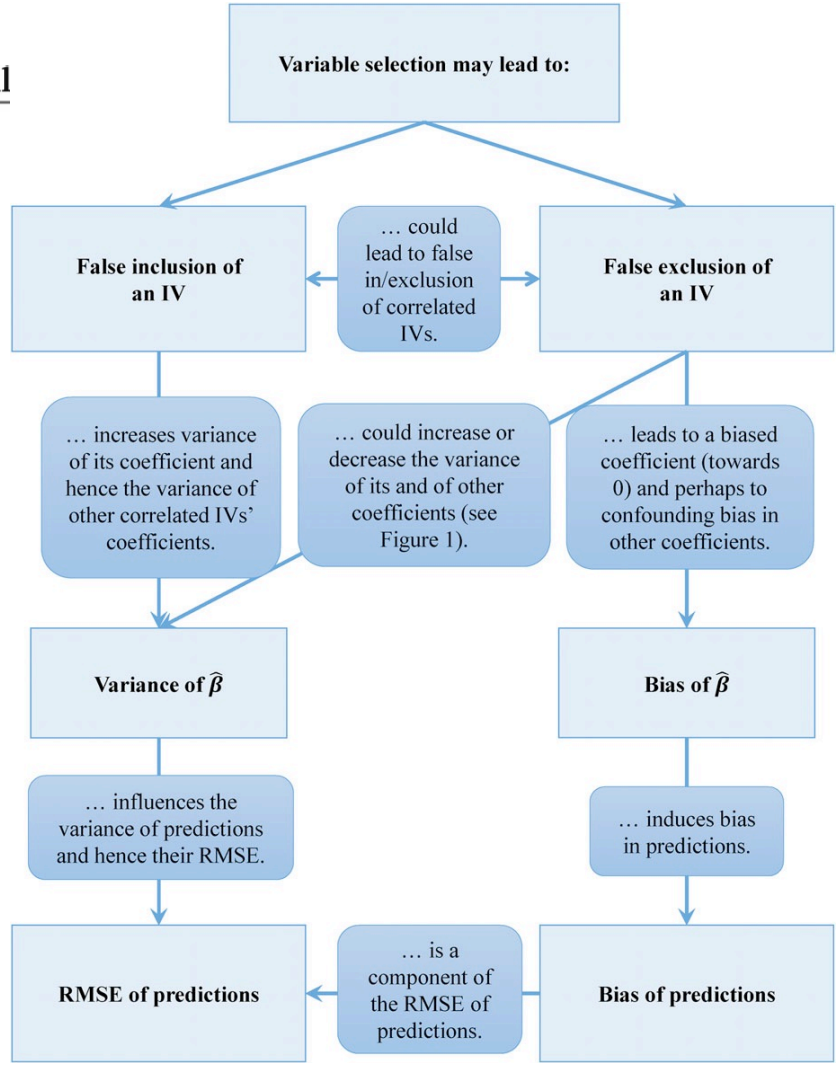
REVIEW ARTICLE

Biometrical Journal

Variable selection – A review and recommendations for the practicing statistician

Georg Heinze | Christine Wallisch | Daniela Dunkler

- The probability of false selections is quite high (multiplicity, sequential testing, sampling variability, …)
- Simulations and resampling suggest that the ‚true‘ Data Generating Model can hardly be identified.
- NOT clear what is to be recommended.



Recommendations

For a 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. Report instabilities and uncertainties, not just final model!

Heinze, Wallisch & Dunkler, BiomJ, 2018

Reporting

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

	Global model			Selected model						
Predictors	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).

As a minimum, report:

- global (initial) model and
- estimated selection probabilities!

The global model also allows to perform unconditional inference

Workshop

Workshop – variable selection (linear only)

1. Data simulation & trying to recover true set of relevant variables:
   Github repository for simulation code: https://github.com/gitMarcH/IBS_Ghana_StratosTG2
   - Proportion of simulations with true set?
   - ---------- with no missing true predictors?
   - ---------- with no irrelevant variables included?

   → Get some intuition how hard it is to identify true data generating model.

2. Body fat data – linear terms only for now
   This dataset is from a Kaggle competition (https://www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset).
   Data obtained from http://hbiostat.org/data courtesy of the Vanderbilt University Department of Biostatistics.
   - Which variable would you include in a model to predict body fat (variable 'fat')?
   - Discuss: can you trust results?
   - Repeat after resampling and summarise selection proportions.

   You can find a description of the variables here: https://hbiostat.org/data/repo/cbodyfat

Workshop – variable selection (linear only)

Data simulation code
(download / copy from GitHub)

```r
## a few Gaussian random variables
k<-12
n<-500
df<-mvrnorm(n=n,mu=rnorm(k,sd=10),Sigma=rWishart(1,k,diag(k))[,,1])
colnames(df)<-paste(sep="","n",1:k)
df<-as.data.frame(df)

## a Poisson, a gamma and a categorical random variable
df<-df %>%
  dplyr::mutate(
    p1=rpois(n=n,lambda=2),
    e1=rgamma(n=n,rate=1,shape=2),
    c1=sample(size=n,paste(sep="","lvl",1:5),prob=c(0.5,0.25,0.125,0.08,0.045),replace=TRUE)
  )

## generate the response variable
df<-df %>%
  dplyr::mutate(
    y=n1+0.5*n2+0.25*n3+0.1*n4+0.05*p1+0.2*e1+ifelse(c1=="lvl2",5,0)+rnorm(n=n,sd=1)
    # n1, n2, n3, n4, p1, e1 are predictors as is level "lvl12" from c1
    # change the sd of the noise term to see effect on variable selection
  )
```

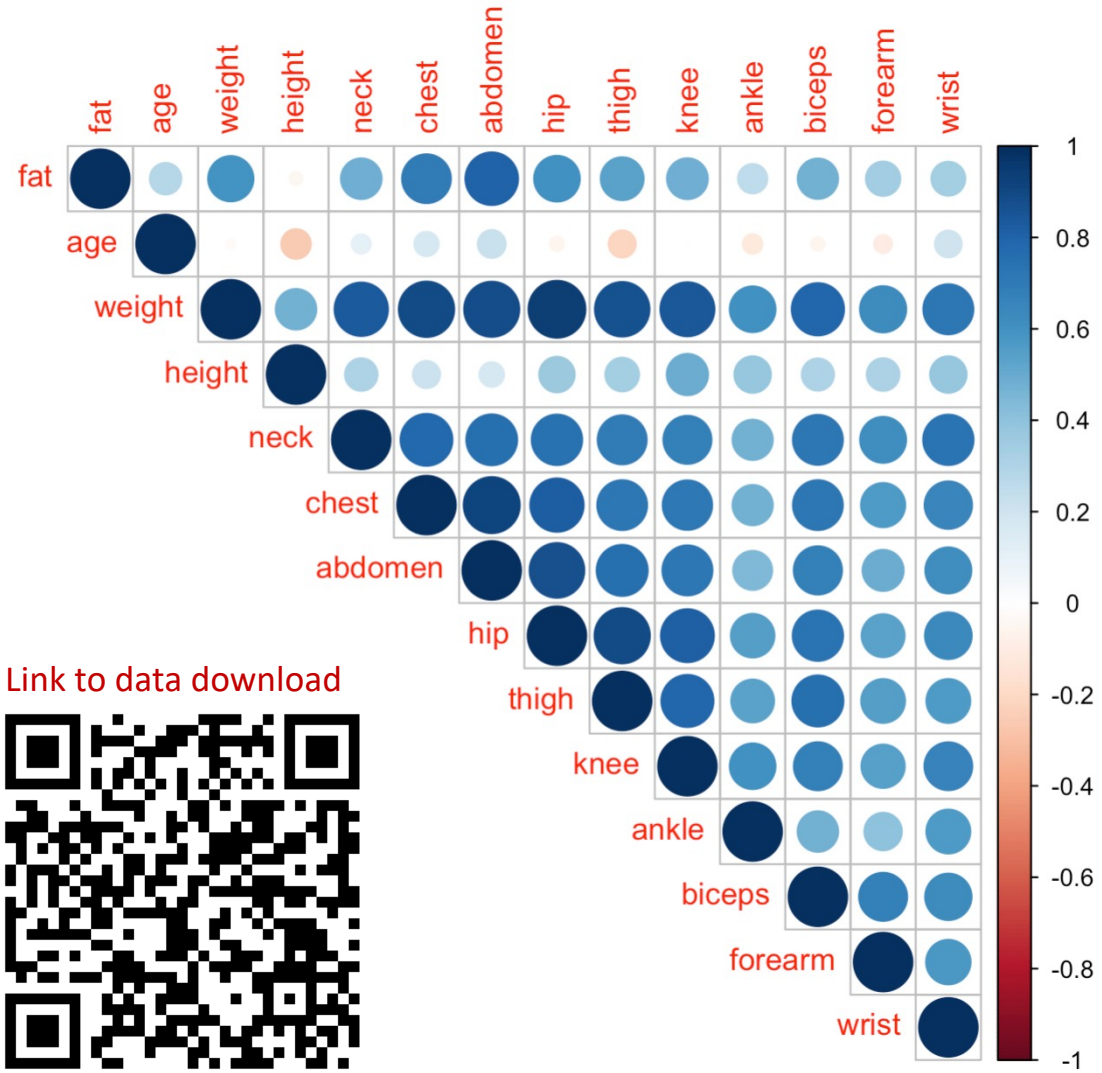
	% correct set	% containing true set	% no irrelevant variables
Univariable	0.0%	0.0%	61.5%
Forward selection	0.0%	19.5%	0.0%
Backward elimination	0.0%	18.0%	0.0%
Lasso	0.0%	0.0%	54.1%

Workshop – variable selection (linear only)

Bodyfat example

Backward selection using AIC (p≈0.157)

```
Call:
lm(formula = fat ~ age + weight + neck + abdomen + hip + thigh +
    forearm + wrist, data = bodyfat, x = TRUE)

Residuals:
Body Fat [Proportion]
      Min        1Q    Median        3Q       Max
-0.110381 -0.030239 -0.001254  0.029603  0.101702

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.2122089  0.1172899  -1.809  0.07166 .
age          0.0005789  0.0003099   1.868  0.06293 .
weight      -0.0019968  0.0008793  -2.271  0.02404 *
neck        -0.0044354  0.0022462  -1.975  0.04946 *
abdomen      0.0094958  0.0007190  13.206  < 2e-16 ***
hip         -0.0019670  0.0014007  -1.404  0.16153
thigh        0.0026583  0.0012982   2.048  0.04168 *
forearm      0.0049547  0.0018601   2.664  0.00825 **
wrist       -0.0150037  0.0050999  -2.942  0.00358 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04265 on 240 degrees of freedom
Multiple R-squared:  0.7383,    Adjusted R-squared:  0.7296
F-statistic: 84.63 on 8 and 240 DF,  p-value: < 2.2e-16
```



Link to data download

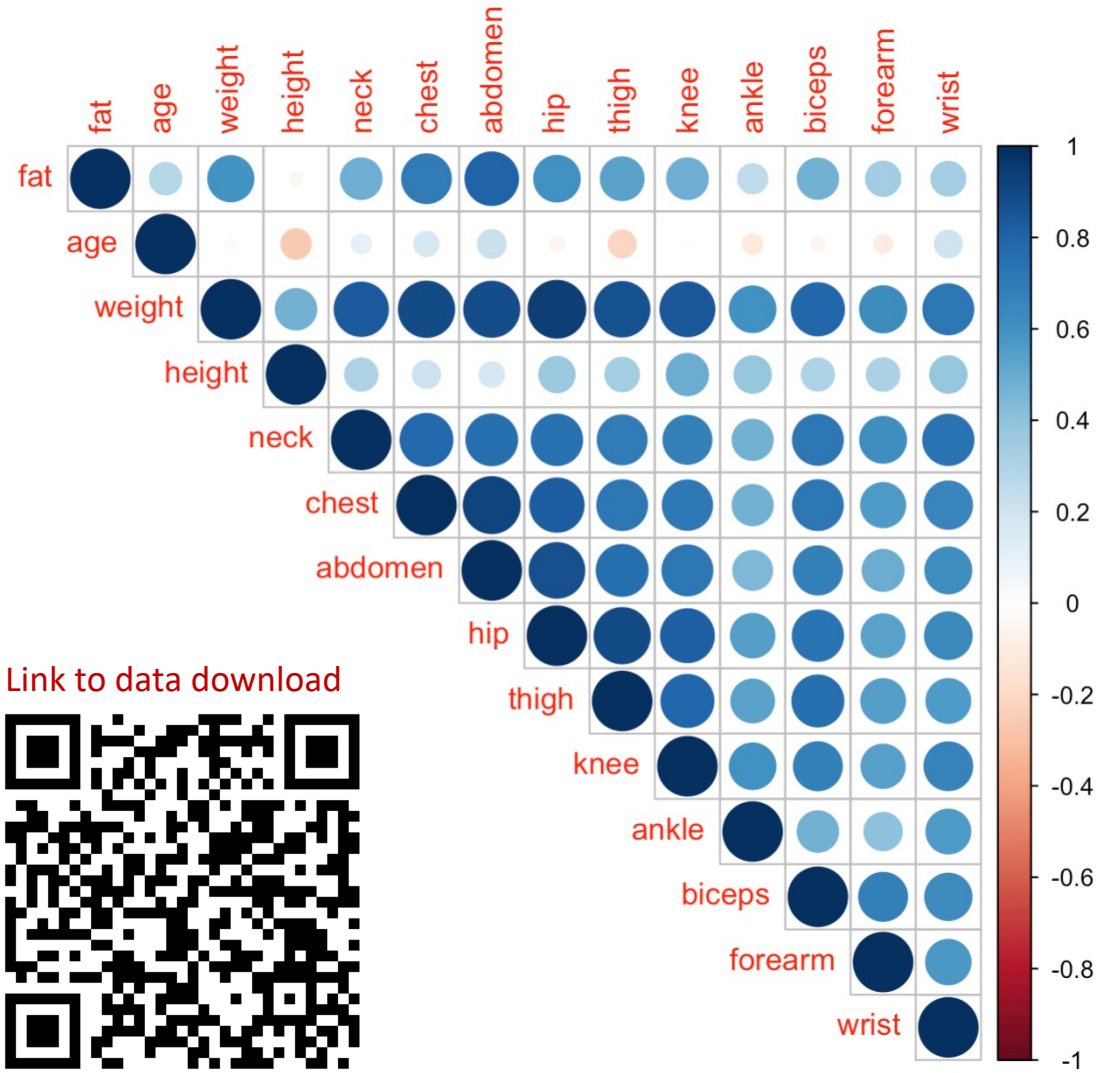
Workshop – variable selection (linear only)

Bodyfat example

Backward selection using AIC (p≈0.157)

Do you trust it?

```
Call:
lm(formula = fat ~ age + weight + neck + abdomen + hip + thigh +
    forearm + wrist, data = bodyfat, x = TRUE)

Residuals:
Body Fat [Proportion]
      Min        1Q    Median        3Q       Max
-0.110381 -0.030239 -0.001254  0.029603  0.101702

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.2122089  0.1172899  -1.809  0.07166 .
age          0.0005789  0.0003099   1.868  0.06293 .
weight      -0.0019968  0.0008793  -2.271  0.02404 *
neck        -0.0044354  0.0022462  -1.975  0.04946 *
abdomen      0.0094958  0.0007190  13.206  < 2e-16 ***
hip         -0.0019670  0.0014007  -1.404  0.16153
thigh        0.0026583  0.0012982   2.048  0.04168 *
forearm      0.0049547  0.0018601   2.664  0.00825 **
wrist       -0.0150037  0.0050999  -2.942  0.00358 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04265 on 240 degrees of freedom
Multiple R-squared:  0.7383,    Adjusted R-squared:  0.7296
F-statistic: 84.63 on 8 and 240 DF,  p-value: < 2.2e-16
```



Link to data download

Workshop – variable selection (linear only)

Bodyfat example

	Full model		Selected model (backward elimination)		Selection probabilities			
	Estimate	SE	Estimate	SE	Univariable	Forward selection	Backward elimination	Lasso
(Intercept)	-0.1533	0.2237	-0.2040	0.1547	100.0% (fixed)	100.0% (fixed)	100.0% (fixed)	100.0 (fixed)%
height	-0.0004	0.0007	-0.0005	0.0006	100.0% (fixed)	100.0% (fixed)	100.0% (fixed)	100.0% (fixed)
abdomen	0.0095	0.0009	0.0102	0.0009	100.0% (fixed)	100.0% (fixed)	100.0% (fixed)	100.0% (fixed)
wrist	-0.0161	0.0053	-0.0204	0.0063	4.7%	91.7%	95.0%	6.9%
forearm	0.0044	0.0020	0.0138	0.0025	5.0%	74.6%	77.2%	12.0%
neck	-0.0046	0.0024	-0.0043	0.0020	4.3%	66.2%	67.0%	7.7%
age	0.0006	0.0003	0.0007	0.0003	4.9%	57.5%	66.2%	13.0%
weight	-0.0019	0.0014	-0.0023	0.0009	4.4%	60.7%	54.8%	0.9%
thigh	0.0021	0.0015			4.6%	39.9%	51.1%	5.7%
hip	-0.0020	0.0015			4.3%	42.9%	49.3%	3.3%
ankle	0.0018	0.0022			5.2%	31.4%	34.1%	11.7%
biceps	0.0016	0.0017			5.1%	34.3%	30.0%	8.8%
chest	-0.0002	0.0010	-0.0015	0.0010	5.2%	27.0%	29.6%	3.8%
knee	0.0002	0.0025			4.1%	17.3%	18.2%	5.0%

Functional form selection

Non-linear associations

- So far, we have assumed linear associations of the independent variables with the outcome.
- Luckily, residual analyses did not give rise to any violations of the linearity assumption
- What if we cannot assume a linear association?

Categorisation of continuous variables

Categorisation of continuous variables often done for 2 reasons:

- “Simpler” interpretation.
- To account for non-linear relationships.

→ polynomial regression, splines and FPs are all better than categorization

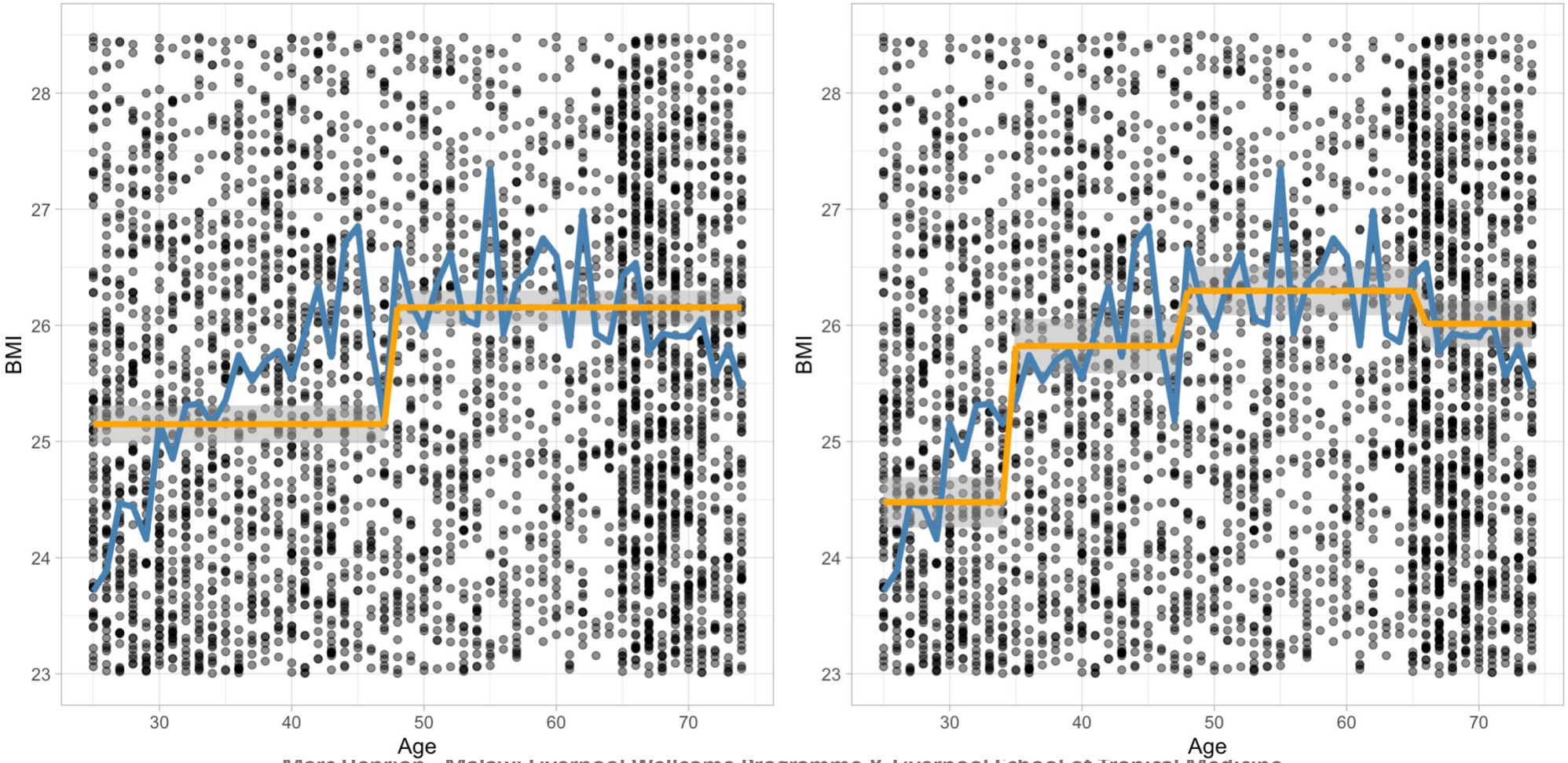
→ interpretation seems at first more difficult, but this is NOT an issue (see next slide…)

Categorisation of continuous variables

- Cut-points are still popular in clinical and epidemiological research.
- Use of cut-points in a model gives a step function.
- Main issues:
  - How many cut-points?
  - Where should the cut-points be put?
  - Biologically implausible step functions are a poor approximation to the true relationship.
  - Almost always fits the data less well than a suitable continuous function.
- Nevertheless, in many areas still the preferred approach!

Categorisation of continuous variables

Piecewise constant fit (orange curve) by year of age with 95% CI (grey band).



Example: NHANES data

- For an example of a descriptive model, we will use one of the NHANES datasets.
- The National Health and Nutrition Examination Survey (**NHANES**) is a survey research program conducted by the National Center for Health Statistics (NCHS) to assess the health and nutritional status of adults and children in the United States, and to track changes over time. (Wikipedia)
- At each 'wave' of the survey, various variables are collected from the survey participants (N~10,000 per wave).

Example: NHANES data

- We are interested in modelling the association of body-mass-index (BMI) with age and gender of the participants.

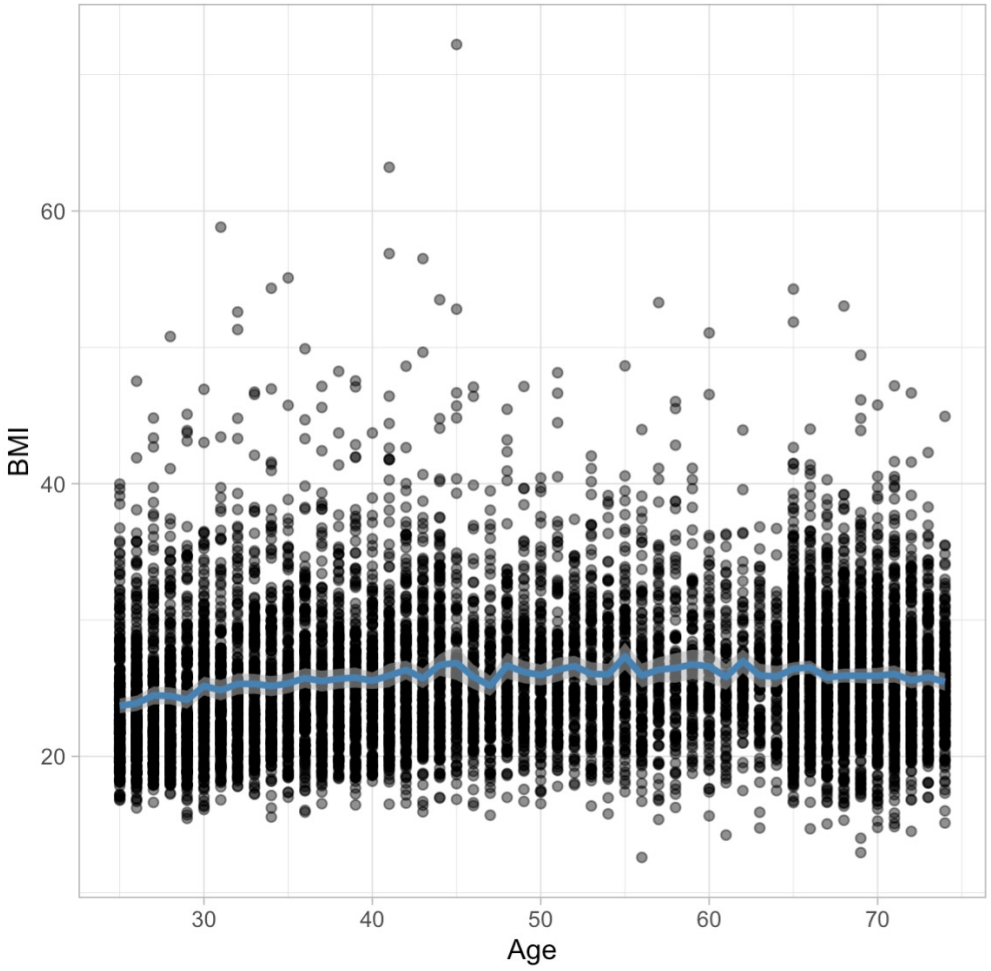
- We use a subset of complete data from one wave of the survey consisting of 8,593 participants (3,901 males [coded 1], 6,031 females [coded 2]), aged 25-74

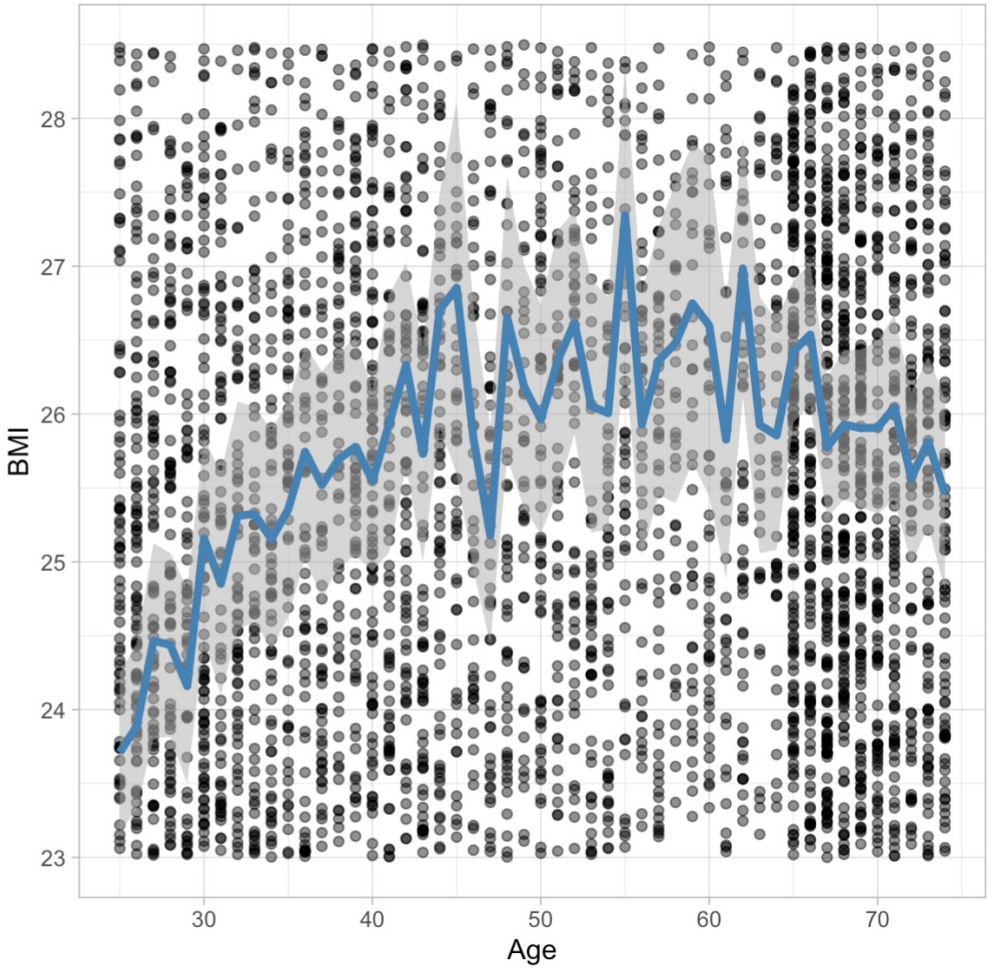
- The median BMI is 24.9 (interquartile range 22.1-28.3, range 12.6-72.2)

https://www.kaggle.com/datasets/homayoonkhadivi/nhanes-datasets

Example: NHANES – BMI as a function of age

Mean BMI (**blue curve**) by year of age with 95% CI (grey band) for means per year of age.





NHANES – descriptive model for BMI

Main question:

How far are differences between adjacent age groups due to a **trend** or due to **random sampling variations**?

→ How to separate systematic changes from unsystematic variation?

We will first try to answer this question for the combined cohort (males & females) and will later separate by sex.

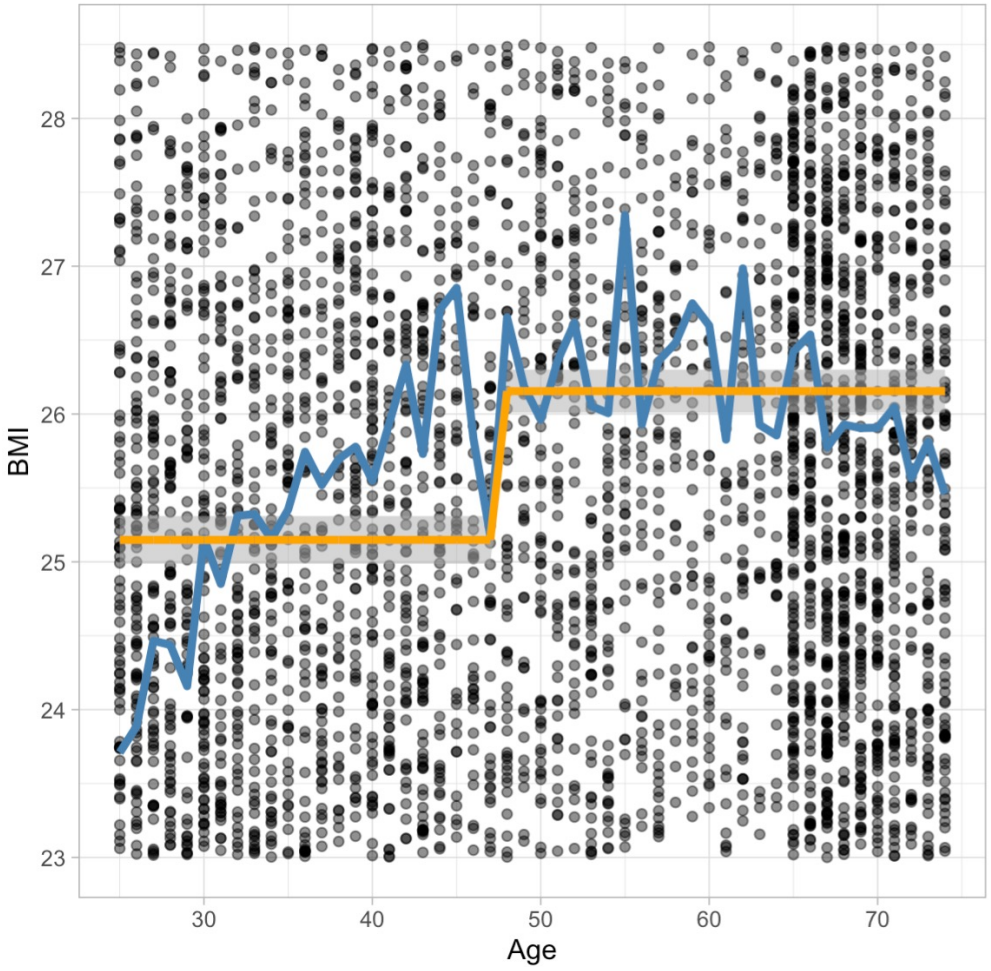
NHANES – descriptive model for BMI

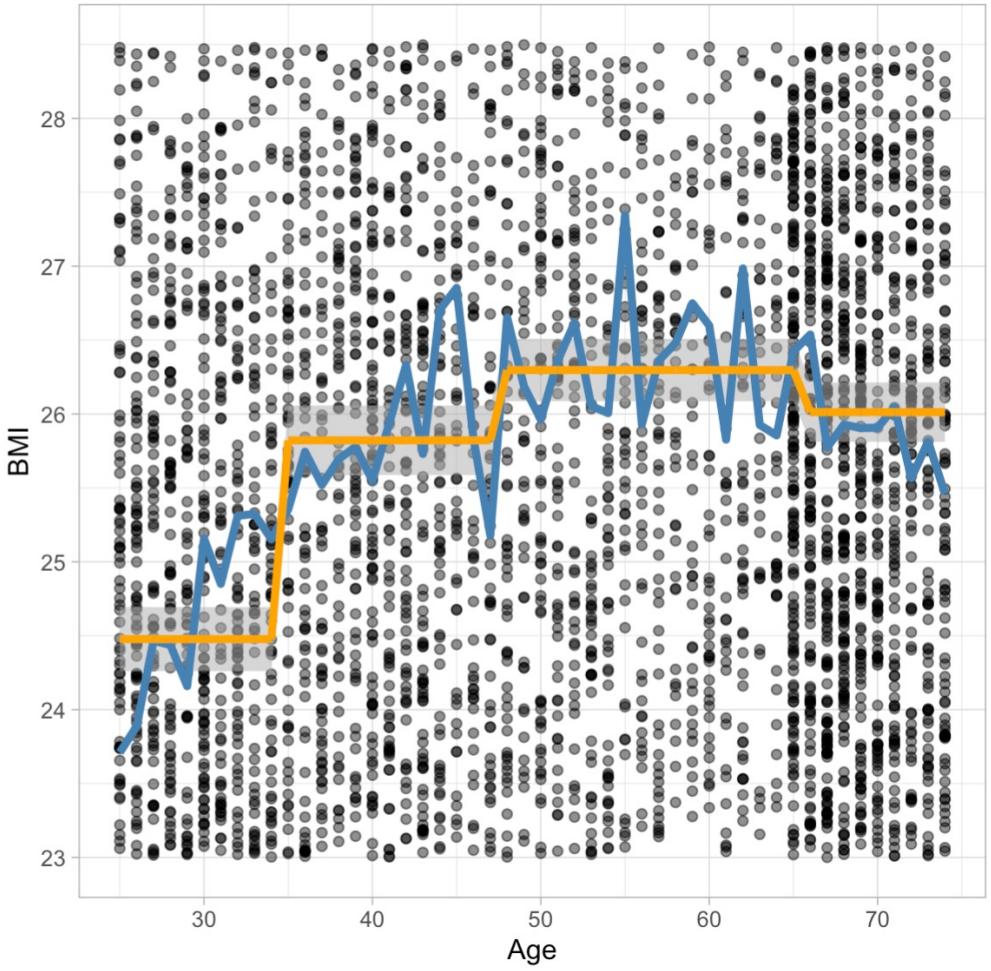
There are many ways we can attempt to capture the non-linear relationship between age and BMI:

- Piecewise constant fit
- Polynomial regression
- Splines
- Fractional polynomials
- Other smoothers

Piecewise constant regression

Piecewise constant fit (orange curve) by year of age with 95% CI (grey band).





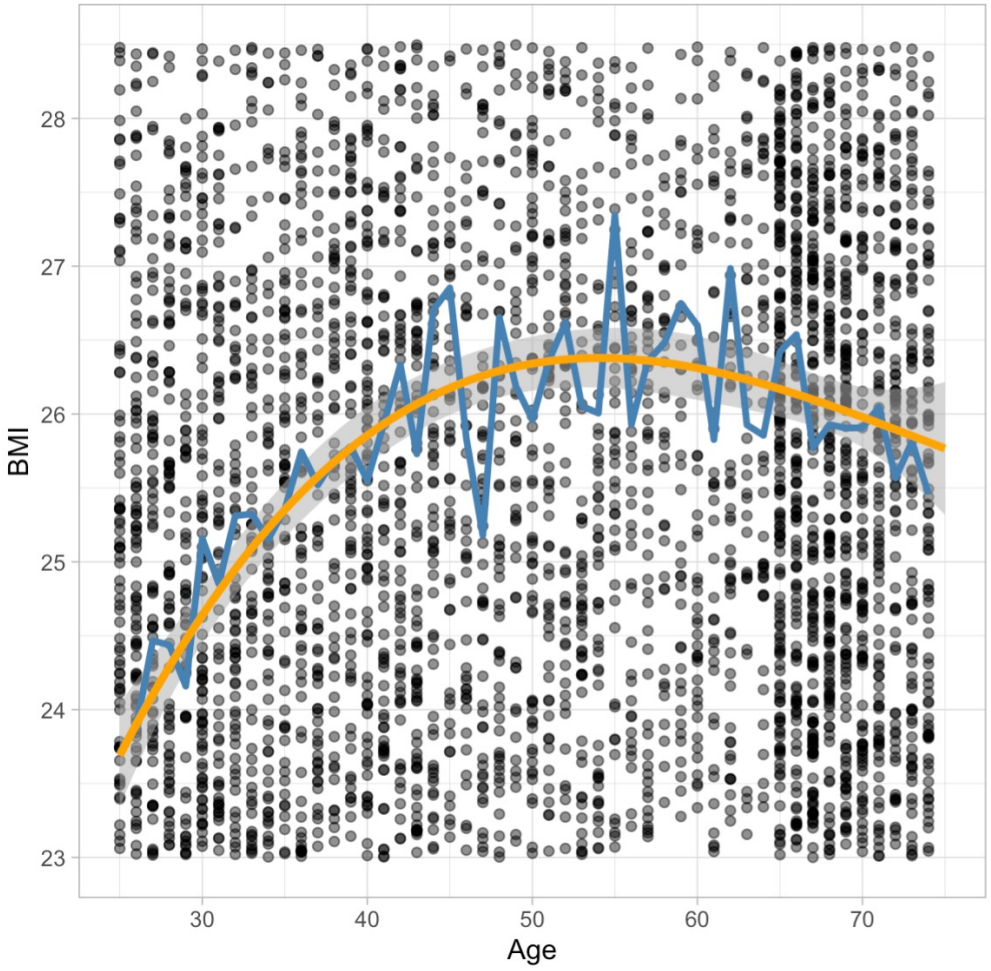
We saw this earlier: this is in effect categorizing a continuous variable.

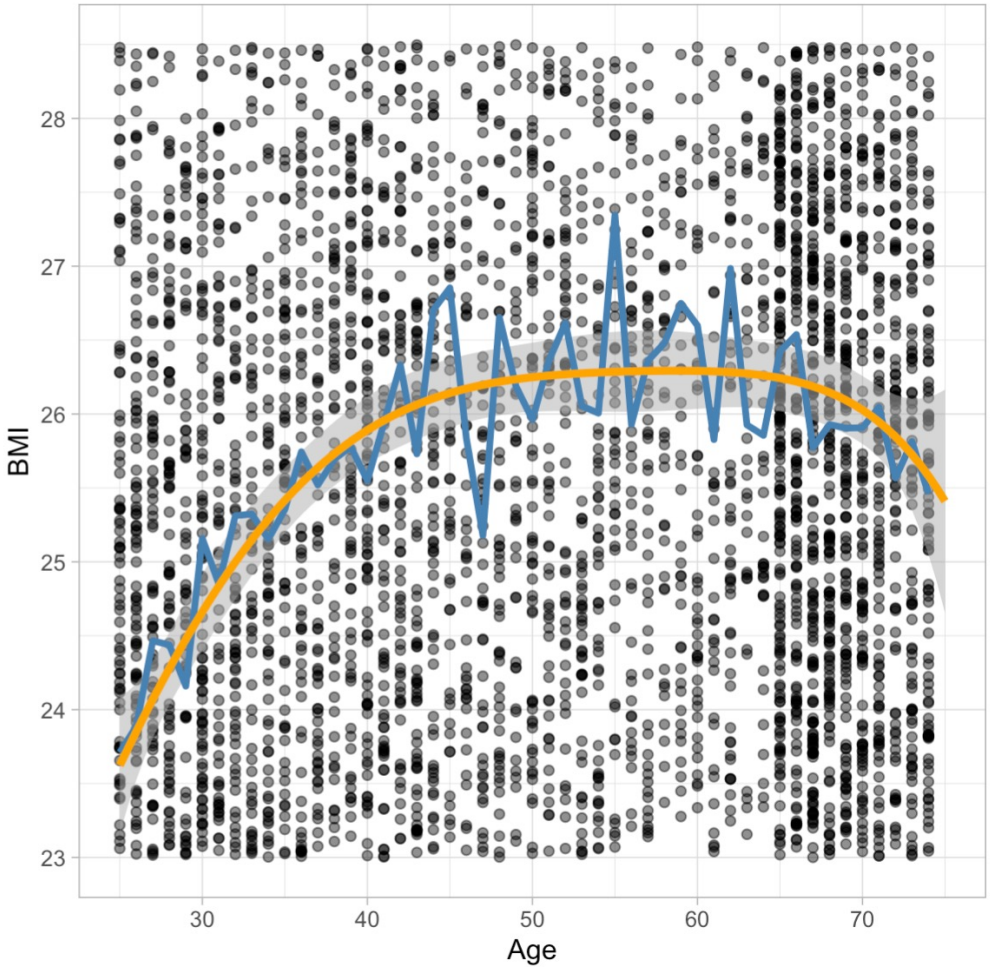
Split predictor range into 2 or more intervals.

Fit an intercept-only model within each interval.

Polynomial regression

Polynomial regression (orange curves) by year of age with 95% CI (grey band); degree 3 / 5 on left/right





Need high degree polynomial for truly flexible shapes.

Non-locality: fit at a point x_0 depends on data far from that point (global not local model).

Create Age^2, Age^3: additional basis functions to age

Fit linear model with Age, Age^2, Age^3 as independent variables

Splines – theory

- Draftman's splines
- General smoothing technique
- Generate basis functions
- Model: fit of y on the basis functions
- **<u>Local</u>** regression → knots
- Many flavours of splines: linear, cubic, B-, natural splines, …

Perperoglou *et al. BMC Medical Research Methodology* (2019) 19:46
https://doi.org/10.1186/s12874-019-0666-3

BMC Medical Research Methodology

REVIEW Open Access

A review of spline function procedures in R

Aris Perperoglou[1*], Willi Sauerbrei[2], Michal Abrahamowicz[3], Matthias Schmid[4] on behalf of TG2 of the STRATOS initiative

- Choices:
  - Basis function set
  - Degree of basis polynomials
  - **Number of knots**
  - **Location of knots**

- Regression & smoothing splines:
  - No penalty → regression splines
  - Smoothing penalty → smoothing splines

Splines – linear splines

Mathematically, for a response y and a predictor x, a linear spline with 3 knots at locations $x = k_1, k_2, k_3$ can be written down as:

$$y = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot (x - k_1)_+ + \beta_3 \cdot (x - k_2)_+ + \beta_4 \cdot (x - k_3)_+ + \epsilon$$

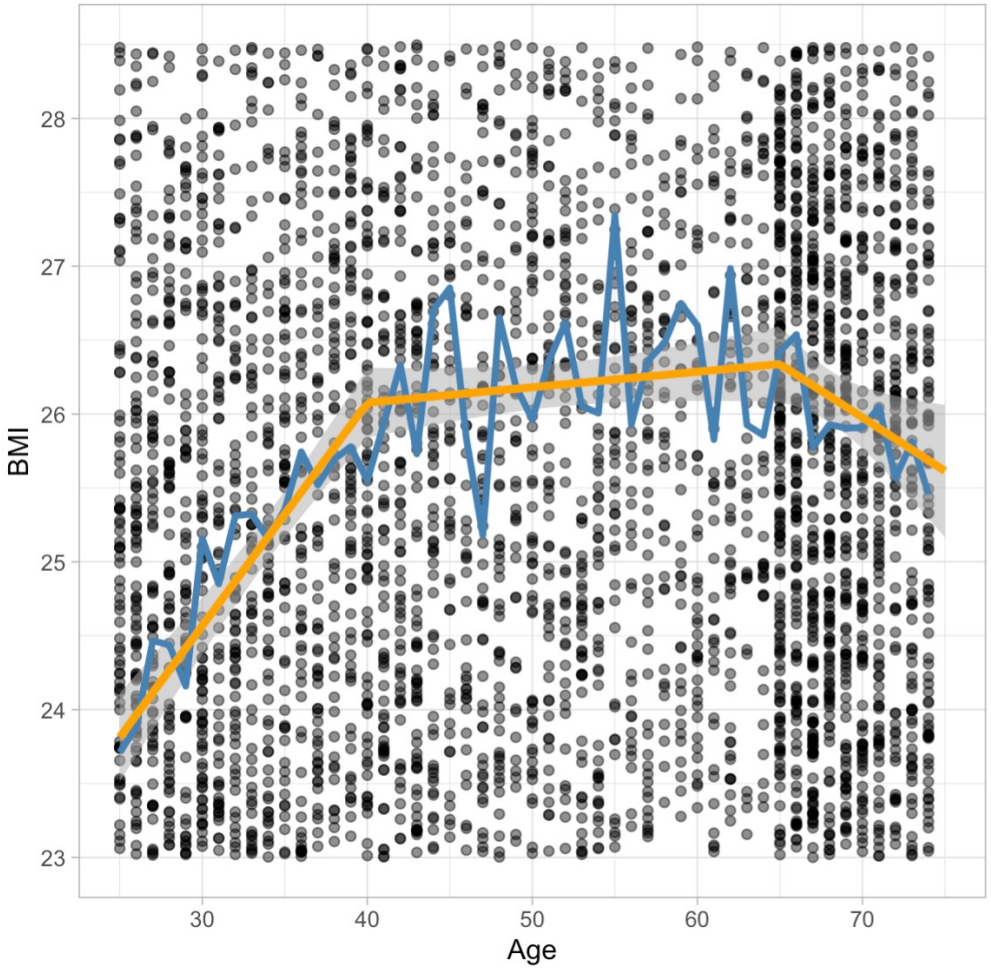
where $(x)_+ = x$ if $x \geq 0$ and 0 otherwise.

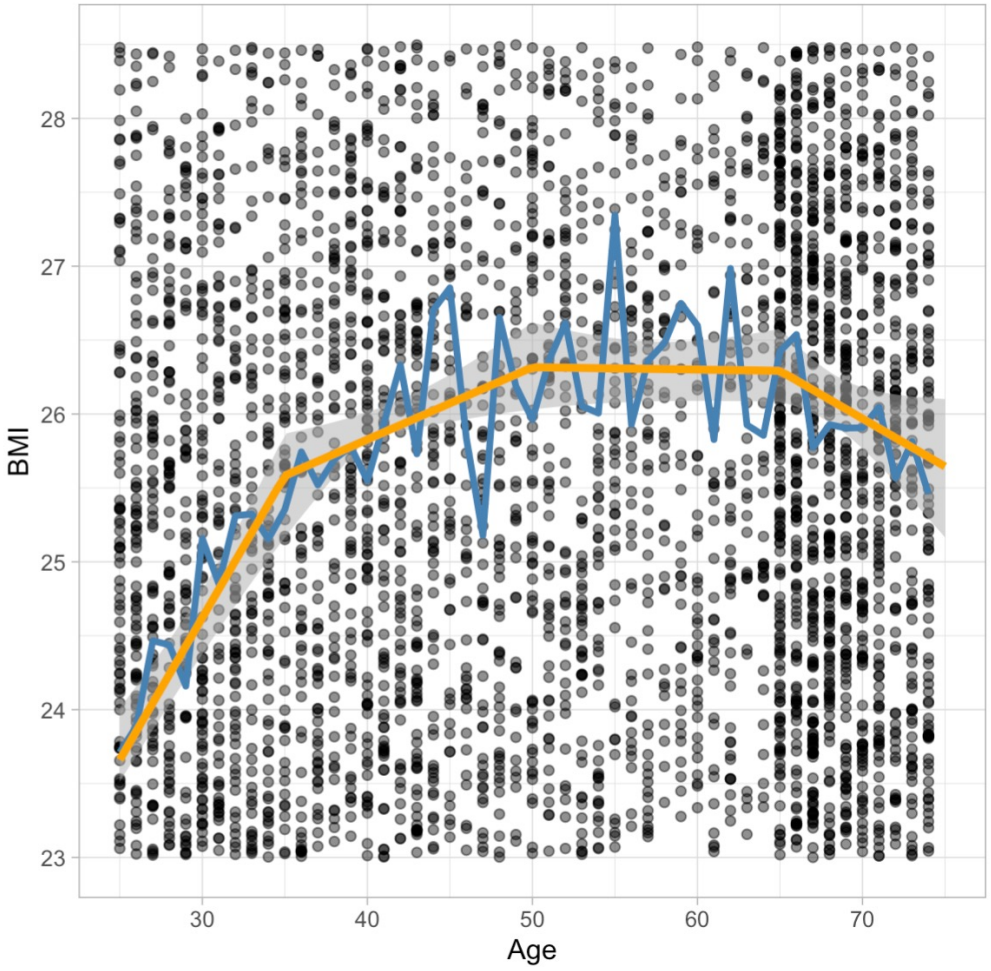
This means that the fitted curve has different slopes in different intervals of values of x:

interval	slope
(-Inf, k1)	β_1
[k1,k2)	$\beta_1 + \beta_2$
[k2,k3)	$\beta_1 + \beta_2 + \beta_3$
[k3,Inf)	$\beta_1 + \beta_2 + \beta_3 + \beta_4$

Splines – linear spline fits

Linear spline regression (**orange curves**) by year of age with 95% CI (grey band); 2/3 knots on left/right

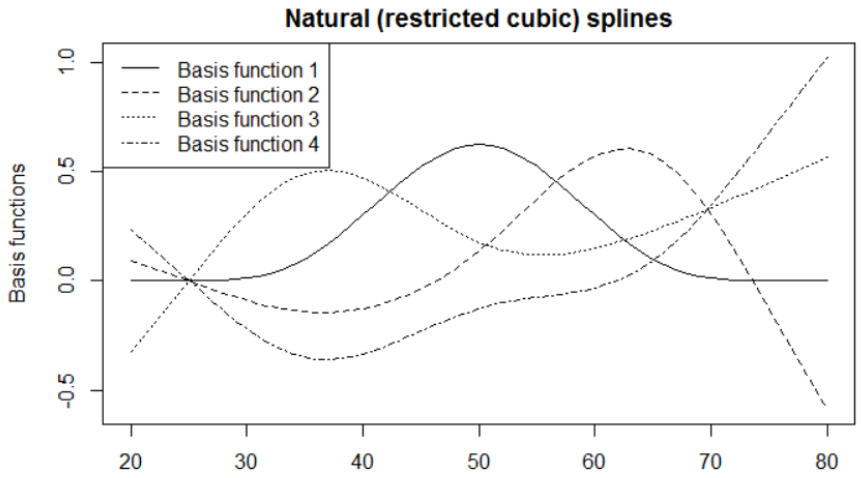




Splines – natural / restricted cubic splines

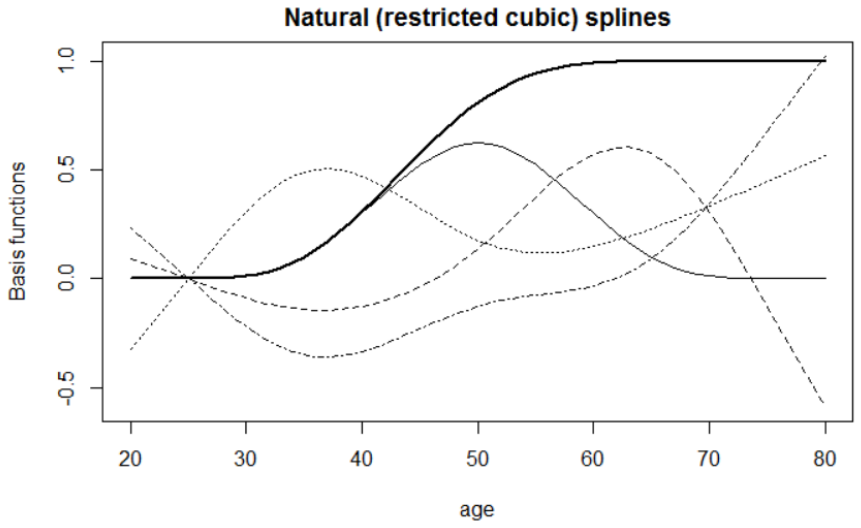
- Specify K knots:
  - Boundary knots
  - Interior knots
- Number of regression coefficients to estimate = K - 1
- Degrees of freedom = K + 1
- Natural splines use:
  - Cubic basis functions.
  - Linear basis functions at the tails. (left and right of boundary knots)
  - They are also called *restricted cubic splines*.

Boundary knots 25, 75; 3 interior knots: 35, 50, 65:



The **unweighted sum of the basis functions** (thick line) gives a sigmoid curve.

By a weighted sum we can get very flexible response functions.



Splines – practical considerations

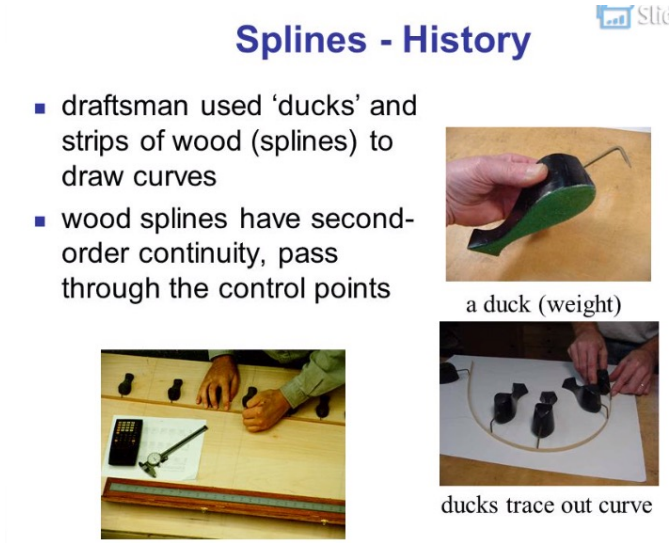
- Some eminent researchers said
  ‘No phenomenon in medicine needs more than 3 degrees of freedom to describe.’

- In this sense, a natural spline with 3 interior and 2 boundary knots should be sufficient for most applications.

- The placement of knots could be guided by
  - Data density (e.g. equidistant quantiles).
  - Desired flexibility in certain regions (e.g. more knots in regions of interest).

- The number of knots can be guided by
  - Common sense (enough flexibility with 3 degrees of freedom?)
  - Information criteria (e.g. Akaike information criterion [AIC], adjusted R^2)

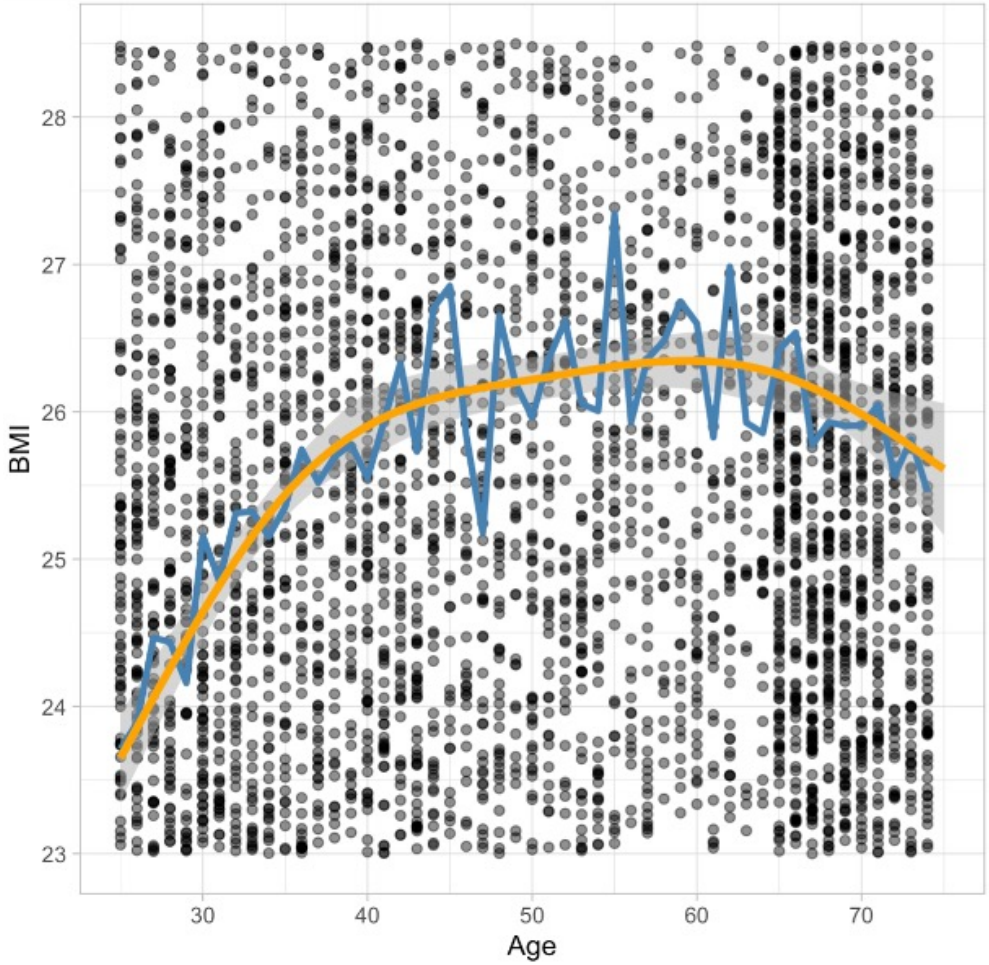
- **Recommendation: explicitly specify the knots and do not use default values!**
  (Transparency, reproducibility!)



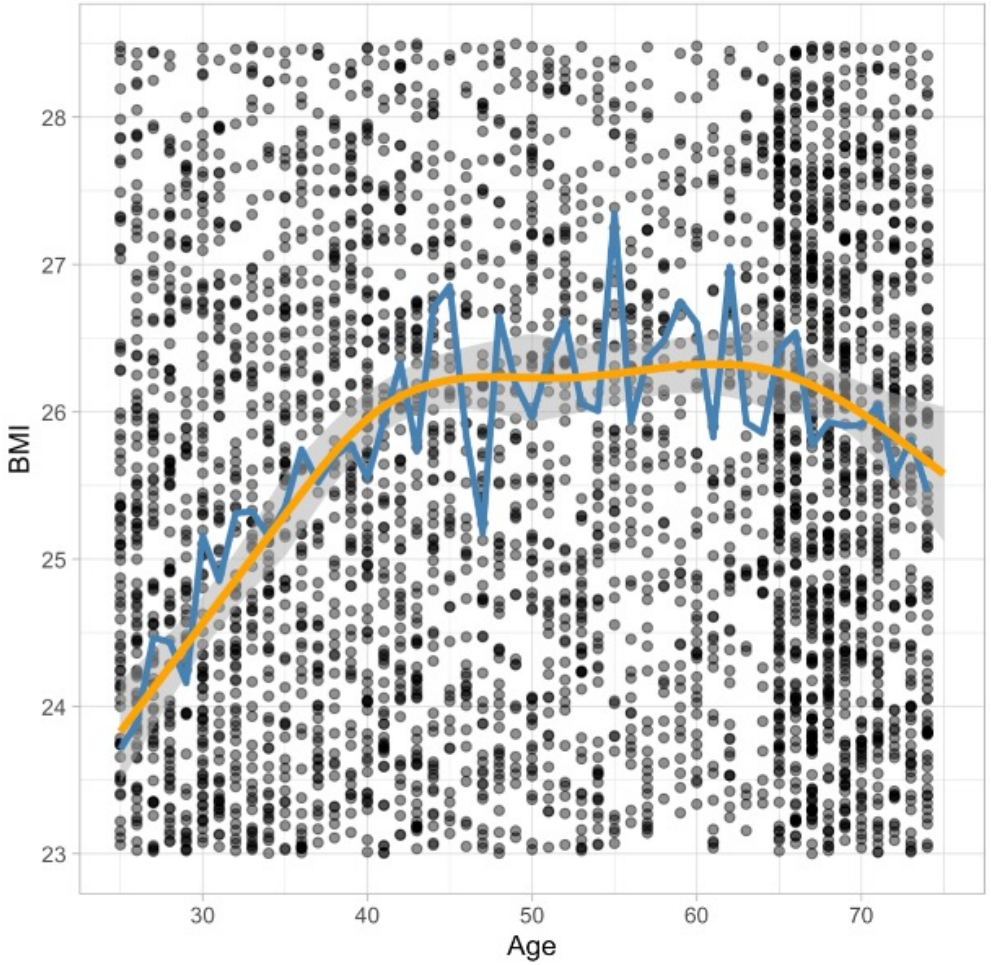
https://slideplayer.com/slide/5191319/

Splines – natural spline fits

Boundary knots (25,75), inner knots (35,50,65)
AIC 52,378; adjusted R^2 0.0218



Boundary knots (35,75), inner knots (42,52,75)
AIC 52,381; adjusted R^2 0.0215



Natural spline regression (orange curves) by year of age with 95% CI (grey band)

Fractional polynomials - theory

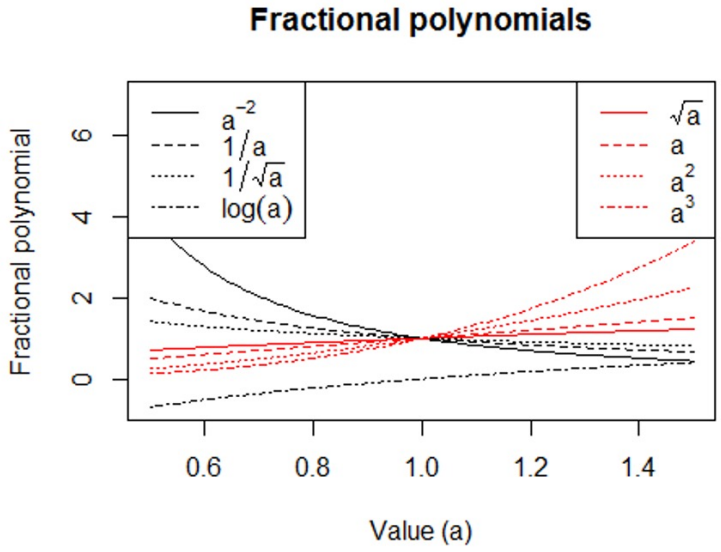
- Select one (FP1) or two (FP2) basis functions out of 8 possible transformations:

$$a^{-2}, a^{-1}, a^{-1/2}, log(a), a^{1/2}, a, a^2, a^3$$

- These basis functions are 'most different' around 1

- Therefore, variables are 'pretransformed' before application:
  - If negative values occur, shift to positive
  - Divide by a multiple of 10 such that SD ~ 1 (not 0.1, not 10)

→ mfp R package does this for you!



Royston and Altman, 1994
Royston and Sauerbrei, 2008

Fractional polynomials – practice

- 'Repeated powers': one can select two equal powers (e.g. 2 and 2):
  In this case the second transformation is multiplied by $\log(a)$, allowing more flexibilty

$$FP2(a, 2,2) = a^2,\ a^2 \log(a)$$

- Selecting the powers (RA2 algorithm):
  - Compute all possible FP2 and FP1 models
  - Start with 'best fitting' FP2 model
    - Compare to linear model – if not significant STOP: linear model selected
    - Compare to best fitting FP1 model – if not significant STOP: FP1 model selected
  - Else: FP2 model selected

- R package: mfp (new package to be released soon)

Fractional polynomials – pros & cons

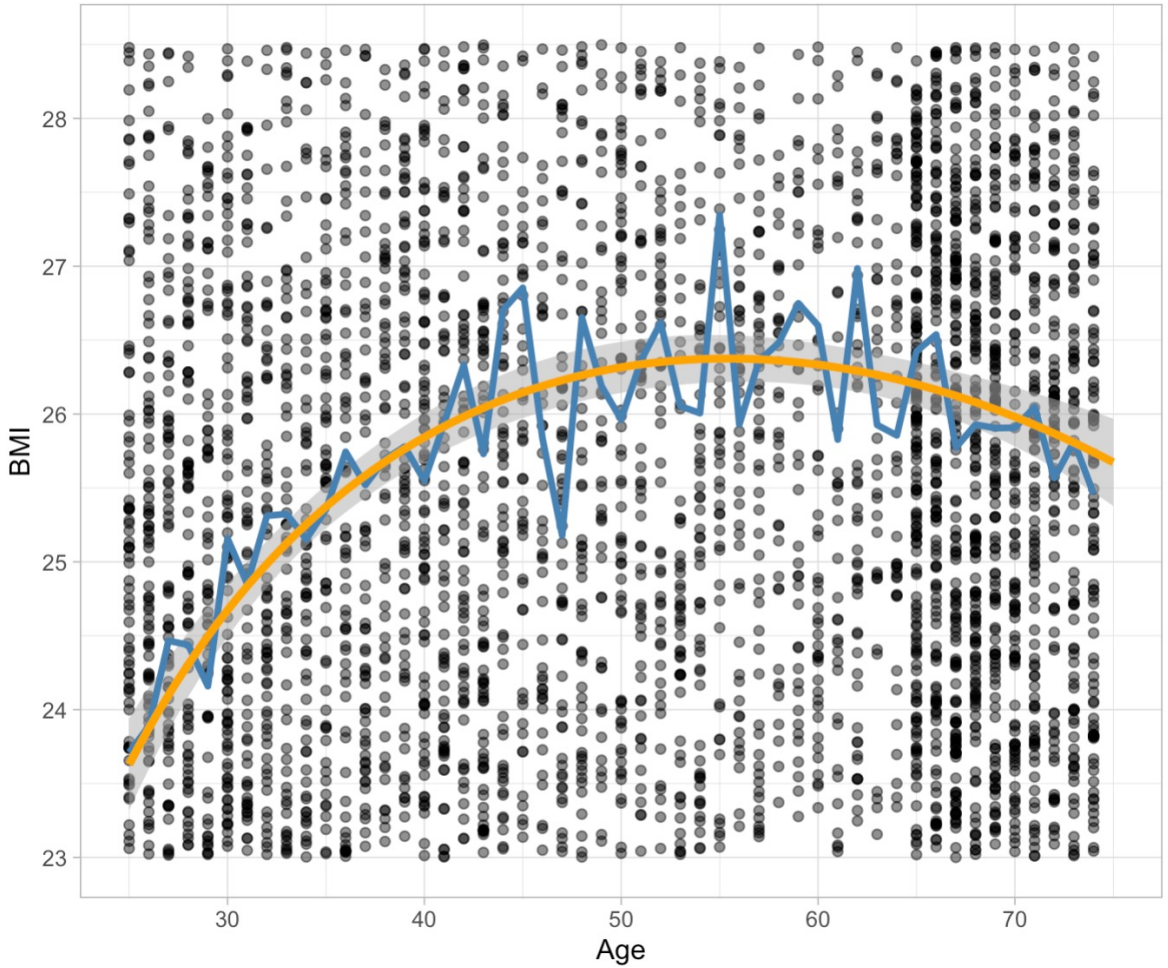
- PRO:
  - Often flexible enough fit for medical applications.
  - Model can be described with max. 2 parameters – very concise.

- CON: ‘optimally selected’
  - Cannot easily model special shapes (e.g. sigmoid curve).
  - Danger of overestimation in small samples.
  - Danger of p-values biased low.

- Selection algorithm (comparison with simpler models) should guard against such problems.

Fractional polynomials – fit

FP regression (**orange curves**) by year of age with 95% CI (grey band)

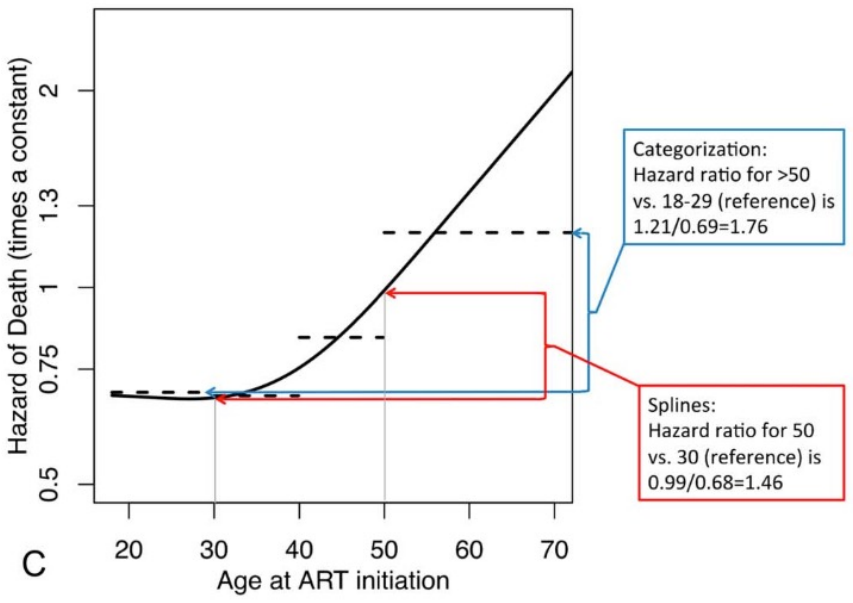


AIC 52,375
Adjusted R^2 0.0219

Interpreting models with splines / FPs

```
Call:
lm(formula = bmi ~ ns(age, knots = c(35, 50, 65), Boundary.knots = c(25,
    75)), data = nhanes_BP)

Residuals:
    Min      1Q  Median      3Q     Max
-15.342  -3.962  -0.692   3.102  47.901

Coefficients:
                                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                                               26.6795     0.1287 207.306  < 2e-16
ns(age, knots = c(35, 50, 65), Boundary.knots = c(25, 75))1   1.3320     0.2991   4.453 8.58e-06
ns(age, knots = c(35, 50, 65), Boundary.knots = c(25, 75))2   3.0082     0.2332  12.899  < 2e-16
ns(age, knots = c(35, 50, 65), Boundary.knots = c(25, 75))3   4.2964     0.3386  12.687  < 2e-16
ns(age, knots = c(35, 50, 65), Boundary.knots = c(25, 75))4  -0.1211     0.1795  -0.675      0.5

(Intercept)                                             ***
ns(age, knots = c(35, 50, 65), Boundary.knots = c(25, 75))1 ***
ns(age, knots = c(35, 50, 65), Boundary.knots = c(25, 75))2 ***
ns(age, knots = c(35, 50, 65), Boundary.knots = c(25, 75))3 ***
ns(age, knots = c(35, 50, 65), Boundary.knots = c(25, 75))4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.714 on 9372 degrees of freedom
Multiple R-squared:  0.03068,	Adjusted R-squared:  0.03027
F-statistic: 74.16 on 4 and 9372 DF,  p-value: < 2.2e-16
```

```
Call:
glm(formula = bmi ~ log((age/100)) + I((age/100)^3), data = nhanes_BP)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-14.981   -3.939   -0.693    3.107   47.821

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     34.2969     0.3911   87.69   <2e-16 ***
log((age/100))   5.3436     0.3296   16.21   <2e-16 ***
I((age/100)^3) -10.9008     0.8387  -13.00   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 32.72564)

    Null deviance: 315682  on 9376  degrees of freedom
Residual deviance: 306770  on 9374  degrees of freedom
AIC: 59324

Number of Fisher Scoring iterations: 2
```



BRIEF REPORT: EPIDEMIOLOGY

Assessing and Interpreting the Association Between Continuous Covariates and Outcomes in Observational Studies of HIV Using Splines

Bryan E. Shepherd, PhD, and Peter F. Rebeiro, PhD,† the Caribbean, Central and South America network for HIV epidemiology*

https://doi.org/10.1097/qai.0000000000001221

Workshop

Workshop

- Fit different types of splines, FPs to NHANES dataset:
  - Try different flavours of splines.
  - Try different numbers and locations of knots.
  - Try fitting different FP models (using mfp() or by hand).
  - Repeat stratified by sex.

- Try interpreting the results.
  Consult Shepherd & Ribeiro (2017) example as a tutorial if needed.

- Re-visit the bodyfat data: what model would you suggest if you want to predict bodyfat from BMI?

Wrap-up & future STRATOS TG2 projects

Wrap-up

Towards state of the art—research required!

Sauerbrei *et al. Diagnostic and Prognostic Research* (2020) 4:3
https://doi.org/10.1186/s41512-020-00074-3

Diagnostic and Prognostic Research

COMMENTARY Open Access

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

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

Table 1 Relevant issues in deriving evidence-supported state of the art guidance for multivariable modelling

No.	Item
1	Investigation and comparison of the properties of variable selection strategies
2	Comparison of spline procedures in both univariable and multivariable contexts
3	How to model one or more variables with a 'spike-at-zero'?
4	Comparison of multivariable procedures for model and function selection
5	Role of shrinkage to correct for bias introduced by data-dependent modelling
6	Evaluation of new approaches for post-selection inference
7	Adaption of procedures for very large sample sizes needed?

Many thanks to all STRATOS TG2 members

STRATOS TG2

G. Heinze (co-chair)
W. Sauerbrei (co-chair)
A. Perperoglou (co-chair)
D. Dunkler
M. Abrahamowicz
M. Kammer
E. Kipruto
F. Harrell Jr
D. Thompson
M. Henrion
N. Klein
K. Luijken
H. Becher
G. Rauch
P. Royston
H. Binder
C. Wallisch
M. Schmid
G. Buch
T. Prince

References

References

- De Boor, C., 2001. A practical guide to splines: with 32 figures, Rev. ed. ed, Applied mathematical sciences. Springer, New York.
- Dunkler, D., Plischke, M., Leffondré, K., Heinze, G., 2014. Augmented Backward Elimination: A Pragmatic and Purposeful Way to Develop Statistical Models. PLoS ONE 9, e113677. https://doi.org/10.1371/journal.pone.0113677
- Harrell, F.E., 2001. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis, Springer Series in Statistics. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4757-3462-1
- Hastie, T., Tibshirani, R., 1999. Generalized additive models. Chapman & Hall/CRC, Boca Raton, Fla.
- Heinze, G., Wallisch, C., Dunkler, D., 2018. Variable selection - A review and recommendations for the practicing statistician. Biom. J. 60, 431–449. https://doi.org/10.1002/bimj.201700067
- Hernán, M.A., Hsu, J., Healy, B., 2019. A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. CHANCE 32, 42–49. https://doi.org/10.1080/09332480.2019.1579578
- Miller, A., 2002. Subset Selection in Regression, 0 ed. Chapman and Hall/CRC. https://doi.org/10.1201/9781420035933
- Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., Schmid, M., 2019. A review of spline function procedures in R. BMC Med Res Methodol 19, 46. https://doi.org/10.1186/s12874-019-0666-3
- Royston, P., Altman, D.G., 1994. Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. Applied Statistics 43, 429. https://doi.org/10.2307/2986270
- Royston, P., Sauerbrei, W., 2008. Multivariable model-building. A pragmatic approach to regression analysis based on fractional polynomials for continuous variables. Wiley, Chichester.
- Sauerbrei, W., Royston, P., Binder, H., 2007. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Statist. Med. 26, 5512–5528. https://doi.org/10.1002/sim.3148
- Shepherd, B.E., Rebeiro, P.F., the Caribbean, Central and South America network for HIV epidemiology, 2017. Brief Report: Assessing and Interpreting the Association Between Continuous Covariates and Outcomes in Observational Studies of HIV Using Splines. JAIDS Journal of Acquired Immune Deficiency Syndromes 74, e60–e63. https://doi.org/10.1097/QAI.0000000000001221
- Shmueli, G., 2010. To Explain or to Predict? Statist. Sci. 25. https://doi.org/10.1214/10-STS330
- Wood, S.N., 2017. Generalized Additive Models: An Introduction with R, 2nd ed. Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279