Regression without regrets – Initial data analysis to support prediction modelling in observational studies

Georg Heinze[1], Marianne Huebner[2], Mark Baillie[3]

[1]Medical University of Vienna, Vienna, Austria; STRATOS-TG2

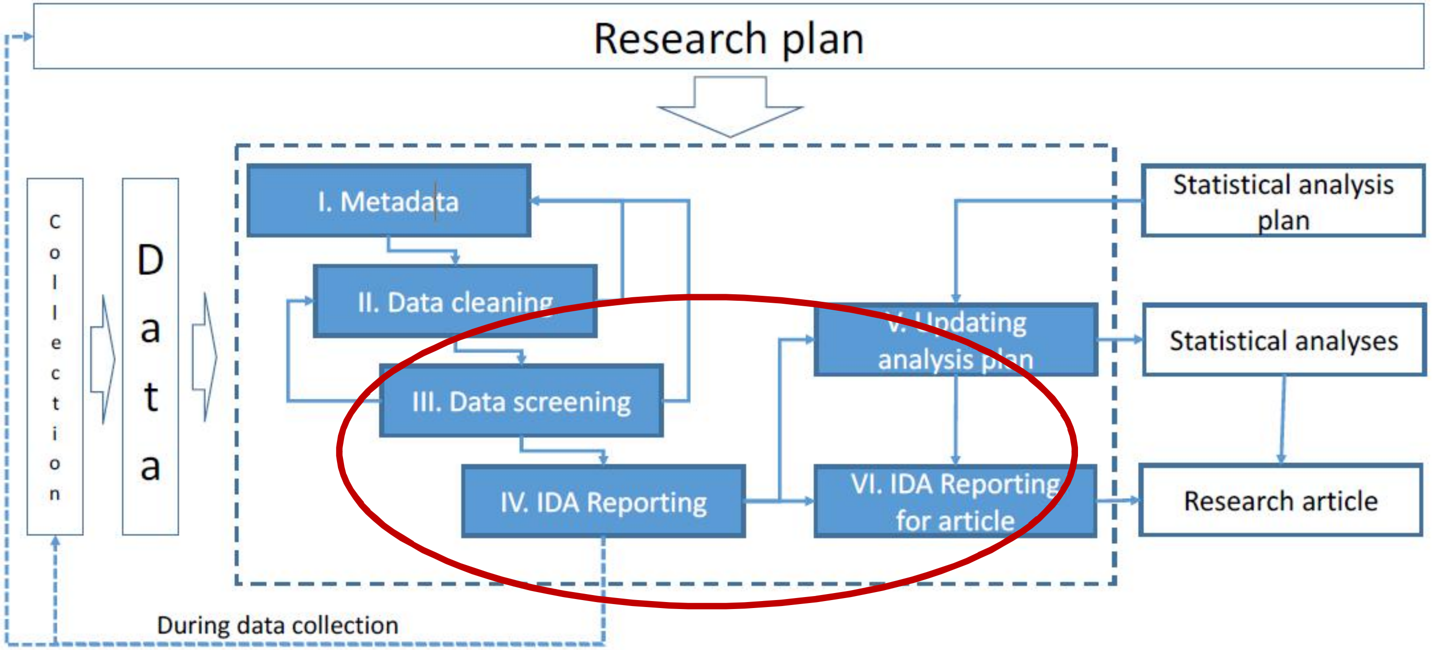
[2]Michigan State University, East Lansing, MI, USA; STRATOS-TG3

[3]Novartis Pharma AG, Basel, Switzerland; STRATOS-TG3, VP

Credits to Sebastian Hödlmoser[1]

What is Initial Data Analysis (IDA)

Huebner et al (2018) defined a framework of IDA consisting of six steps:



Here we concentrate on some aspects relevant for regression modeling.

IDA for regression

- Data screening:
  provide context about data properties and structures
  (to avoid pitfalls)
- In order to:
  - Make decisions about updating statistical analysis plan
  - Help interpreting results of regression analyses
  - Guide presenting results of regression analyses

Initial look into the data,
but not to generate new hypotheses

Do not yet evaluate predictor-outcome associations

IDA ≠ Exploratory data analysis (EDA)

Can IDA be preplanned?

- A statistical analysis plan should contain IDA aspects
- Various steps can be predefined
- Possible consequences can be defined (if IDA results in …, I will do …)

- Aim of this presentation:
  - Describe a generic approach to IDA for regression

Our assumptions and scope

Aspect	Assumptions	Aspect	Assumptions
Purpose of analysis:	Descriptive or predictive model	Type of analysis:	Regression with one outcome variable
Type of outcome:	Continuous, binary or count	Number of predictors:	3-50
Data cleaning:	Completed	Statistical analysis plan:	Exists
Background knowledge:	Collaboration with domain expert		

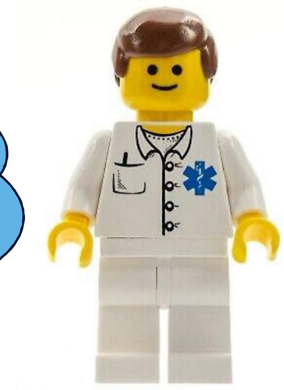
Example study

- Aim: *diagnostic prediction* of bacteremia status with 50 blood analysis predictors + age
- Outcome: binary (bacteremia present or absent)
- Sample size: 14,691
- Source: Clin. Dept. of Laboratory Medicine, Vienna General Hospital
- Availability: data is publicly available
- Publication: Ratzinger F, et al. A Risk Prediction Model for Screening Bacteremic Patients: A Cross Sectional Study. *PLoS ONE* 9(9): e106765. https://doi.org/10.1371/journal.pone.0106765

How the statistical analysis plan is developed

Research aim!!

Have you got data?



Binary outcome (prev. ~10%), 50 continuous predictors, 14,691 observations

Logistic regression with fractional polynomials!

Which are your key predictors?

(searches literature) WBC, NEU, AGE, PLT, BUN, CREA

Any other important (groups of) predictors?

Acute phase reaction indicators
Kidney function indicators

Structural covariates?

AGE, SEX, WBC

Prerequisites

Now I know a lot about bacteremia, but will the data be sufficient?

Can I stick to the plan of using Log Reg with MFP?

How to interpret the model?

How to present the model to my colleague?

Hello statistician! My name is Ida!

Let me show you my IDA checklist.

A (preliminary) checklist for an initial data analysis plan

Topic	Item	Features
Prerequisites		
Statistical analysis plan	P1	Check definition of models and roles of variables in the models
Data dictionary	P2	Check variable labels, definitions, values, units of measurement, type (variables in the SAP)
Domain expertise	P3	Identify groups of predictors, expected proportion of missing values, expected distributions of and correlations between predictors, key predictors, structural covariates for IDA
IDA domain: Missing values (predictor and outcome variables)		
Prevalence	M1	Provide number and proportion of missing values for each predictor, for the outcome variable and for the analysis as a whole; distinguish by type of missingness, if applicable
Patterns	M2	Investigate patterns of missing values across all variables, either as tables or appropriately visualized
IDA domain: Univariate analyses (predictors and outcome)		
Categorical variables	U1	Summarize frequency and proportion for each category or with ordinal plots
Continuous variables	U2	Inspect distributions with high-resolution histogram, summary of main quantiles (e.g. 1st, 5th, 25th, 50th, 75th, 90th, 99th), 5 highest and 5 lowest values, mean, first four moments (mean, variance/standard deviation, skewness, curtosis), number of distinct values. Similarly, inspect distributions of transformed variables, if applicable.
IDA domain: Multivariate analyses (predictors only)		
Correlation	V1	Quantify association with pairwise correlation coefficients between all independent variables in a matrix or heatmap
Association	V2	Visualize the association of each predictor with the structural covariates
Stratification, if applicable	V3	Compute summary statistics for predictors and visualize distributions stratified by structuring covariates
Interactions, if applicable	V4	Evaluate bivariate distributions of the predictors specified in interactions. Include appropriate graphical displays.

Prerequisites

Topic	Item	Features
Prerequisites		
Statistical analysis plan	P1	Check definition of models and roles of variables in the models
Data dictionary	P2	Check variable labels, definitions, values, units of measurement, type (variables in the SAP)
Domain expertise	P3	Identify groups of predictors, expected proportion of missing values, expected distributions of and correlations between predictors, key predictors, structural covariates for IDA

We talked about this before, didn‘t we?

Domain expertise - examples

Some more expertise!

- Dependencies between variables:
  - NEU + BASO + EOS + LYM + MONO ≈ WBC
  - NEUR = NEU / (NEU + BASO + EOS + LYM + MONO) etc.

This must give rise to redundancies...

IDA domain: missing values

IDA domain: Missing values (predictor and outcome variables)		
Prevalence	M1	Provide number and proportion of missing values for each predictor, for the outcome variable and for the analysis as a whole; distinguish by type of missingness, if applicable
Patterns	M2	Investigate patterns of missing values across all variables, either as tables or appropriately visualised
Missing values - Extensions		
Predictors	ME1	Investigate predictors of missingness (complete vs incomplete cases)
Occurrence	ME2	Examine levels of occurrence (within a variable, within an individual, for individuals within strata)

IDA domain: missing values, examples

- Missingness per-variable, per-group, patterns

How much data would be available to fit models with:

Key predictors only?

Key predictors + acute phase + kidney?

All predictors??

Complete cases:

93.9%

63.9%

27.1%

IDA domain: univariate distributions

IDA domain: Univariate analyses (predictors and outcome)		
Categorical variables	U1	Summarize frequency and proportion for each category or with ordinal plots
Continuous variables	U2	Inspect distributions with high-resolution histogram, summary of main quantiles (e.g. 1st, 5th, 25th, 50th, 75th, 90th, 99th), 5 highest and 5 lowest values, mean, first four moments (mean, variance/standard deviation, skewness, curtosis), number of distinct values. Similarly, inspect distributions of transformed variables, if applicable.
Univariate analyses – Extensions		
Sparsity	UE1	Create interactive plots of distributions to inspect sparse ranges or unexpected values
Levels	UE2	Compute summary statistics and describe variation between levels of measurements, e.g. centers, providers, locations

Univariate distributions

Evaluate/refine exclusion criteria?

Is my model robust against influential points?

Should I winsorize or transform predictors?

7.1.2 Structural variables and key predictors

Structural variables and key predictors

7 Variables **14691 Observations**

WBC: White blood count G/L

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
14229	462	2710	1	11.23	7.602	2.66	4.26	6.63	9.60	13.53	18.22	22.27

lowest : 0.00 0.01 0.02 0.03 0.04 , highest: 365.30 383.74 387.73 433.83 604.47

AGE: Patient Age years

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
14691	0	85	1	56.17	20.78	24	29	43	58	70	79	84

lowest : 16 17 18 19 20 , highest: 96 97 98 99 101

SEX: Patient sex 1=male, 2=female

n	missing	distinct	Info	Mean	Gmd
14691	0	2	0.73	1.419	0.4869

```
Value           1      2
Frequency    8536   6155
Proportion  0.581  0.419
```

BUN: Blood urea nitrogen mg/dl

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
14519	172	947	1	22.66	16.92	7.1	8.6	11.6	16.6	26.9	44.8	60.8

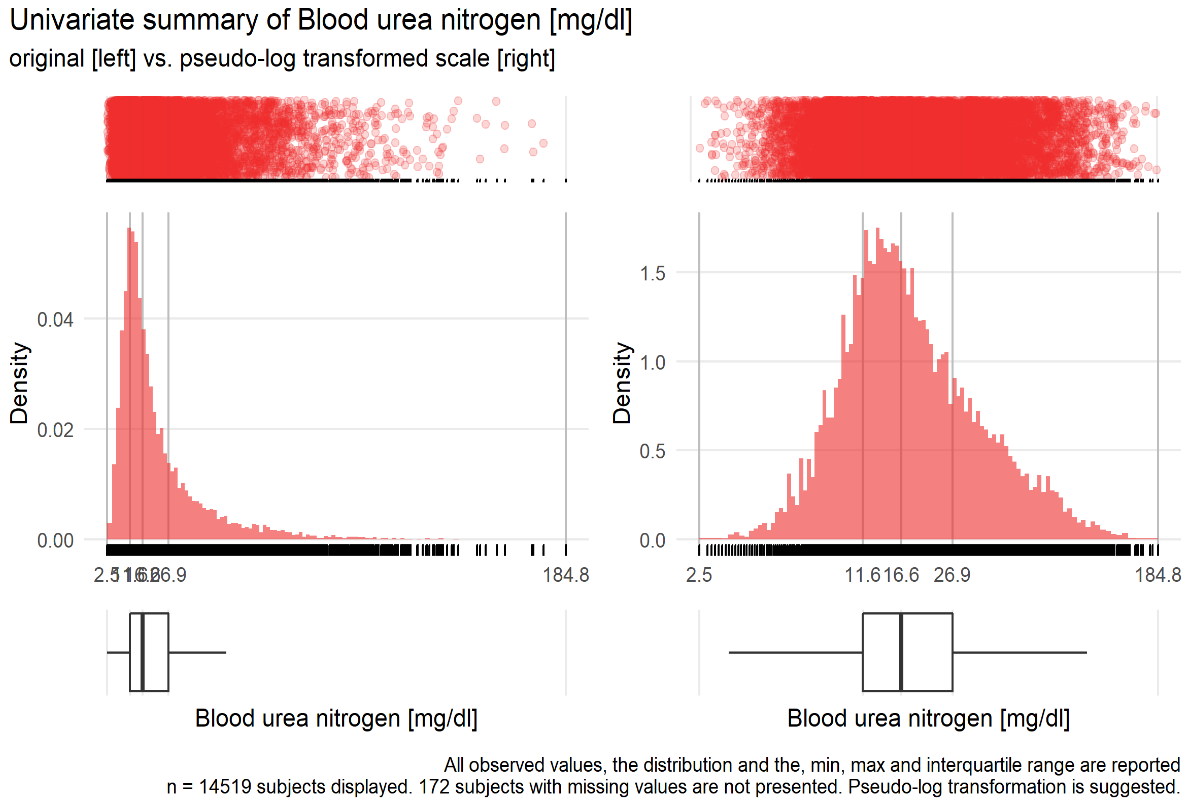
lowest : 2.5 2.7 2.8 2.9 3.0 , highest: 160.6 171.3 171.9 176.0 184.8

MEDICAL UNIVERSITY OF VIENNA

MICHIGAN STATE UNIVERSITY

NOVART

Univariate distributions



A log transformation stabilizes the distribution of this predictor

But it will change the interpretation of the betas!

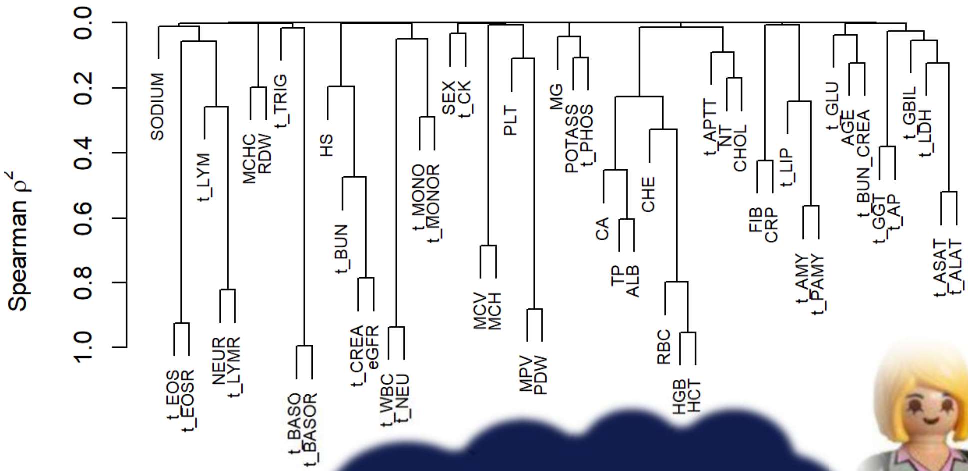
IDA domain: multivariate distributions

IDA domain: Multivariate analyses (predictors only)		
Correlation	V1	Quantify association with pairwise correlation coefficients between all independent variables in a matrix or heatmap
Association	V2	Visualize the association of each predictor with the structuring covariates
Stratification, if applicable	V3	Compute summary statistics for predictors and visualise distributions stratified by structuring covariates
Interactions, if applicable	V4	Evaluate bivariate distributions of the predictors specified in interactions. Include appropriate graphical displays.
Redundancy	V5	Compute Variance Inflation Factors (or multiple R^2 among predictors)
Multivariate analyses – Extensions		
Correlation	VE1	Compare matrix of Spearman and Pearson correlations coefficients
Correlation/ Clustering	VE2	Construct a dendrogram to show closely associated predictors
Redundancy	VE3	Fit parametric additive models to determine how well each predictor can be predicted from the remaining covariates

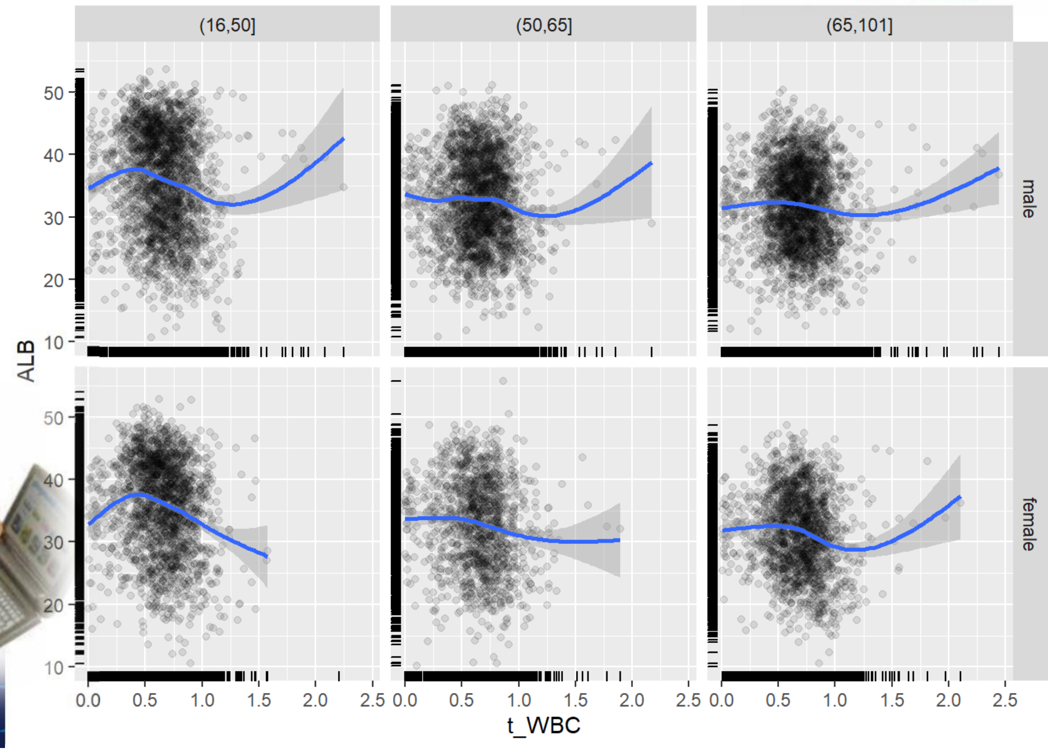
- Here is where the ‚structural covariates' come into play

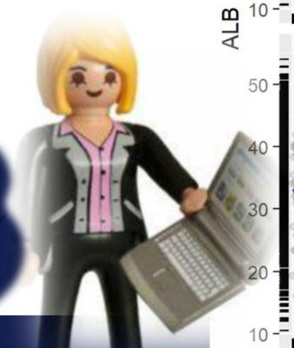
IDA domain: multivariate distributions, examples

- Get an overview: need for structuring (not all scatterplots for any pairs of predictors)
- ALB by WBC (transformed) in six age/sex groups



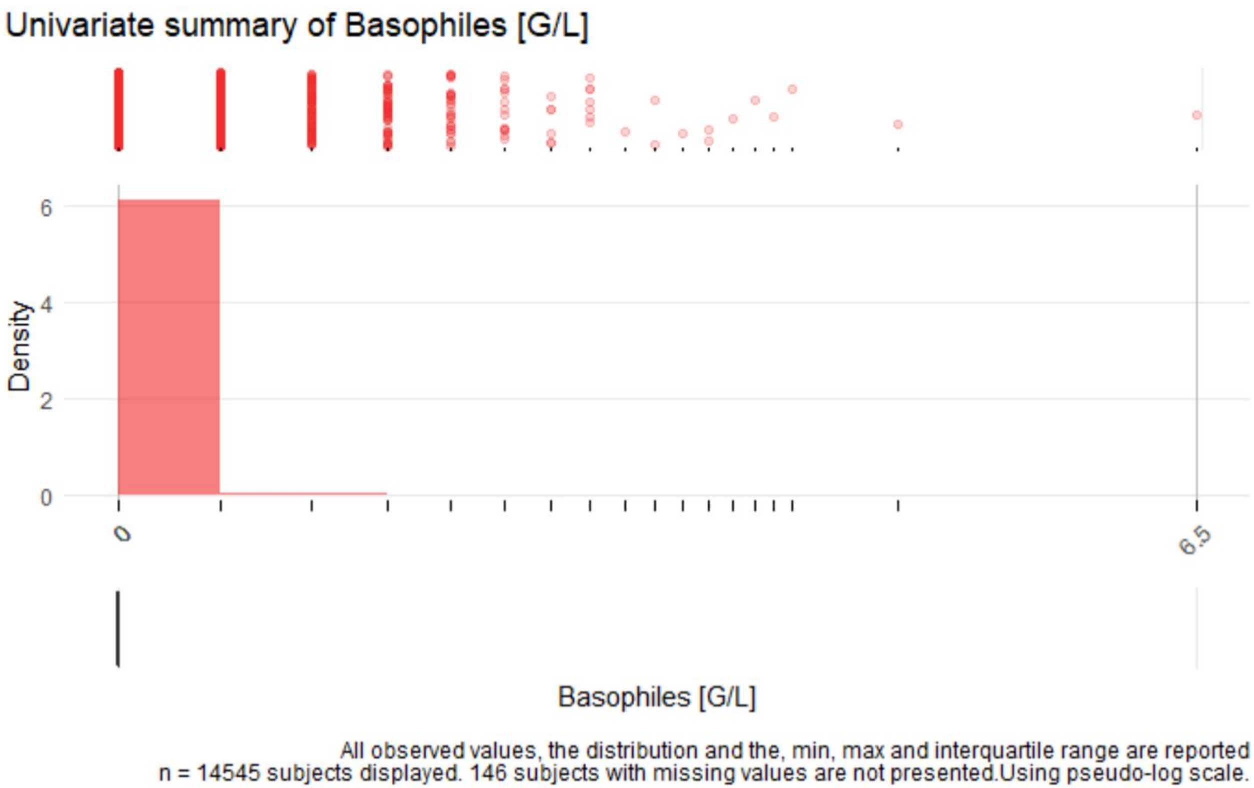
Now we work with some predictors log-transformed (t_XXX)





Really a lot to do here! Good idea to structure that report!

Possible consequences



Transformation does not make that distribution nicer

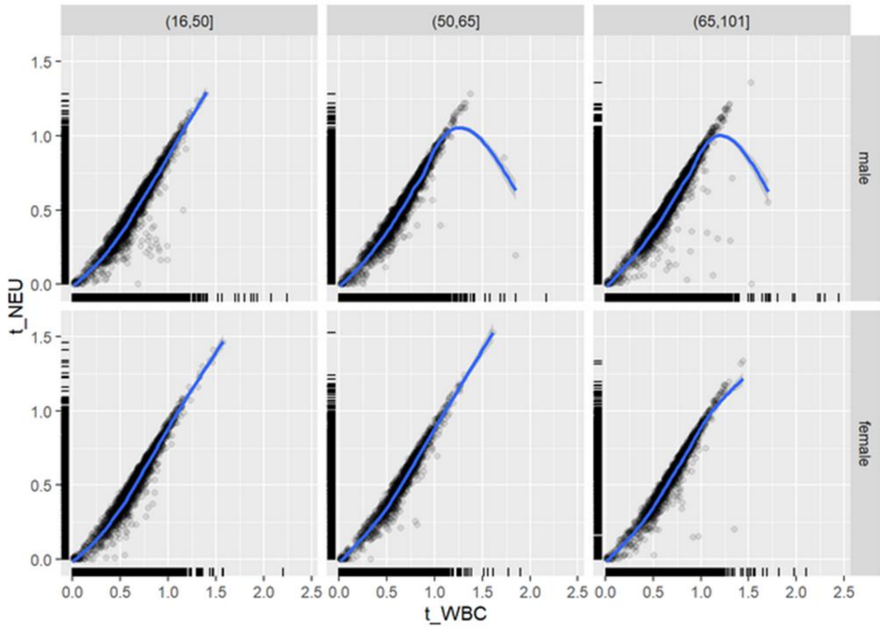
Well, it is what it is.

Possible consequences

You said that there are components of WBC...

Hence, NEU and WBC are indeed highly correlated...

Yes, and NEU is the biggest component of WBC!



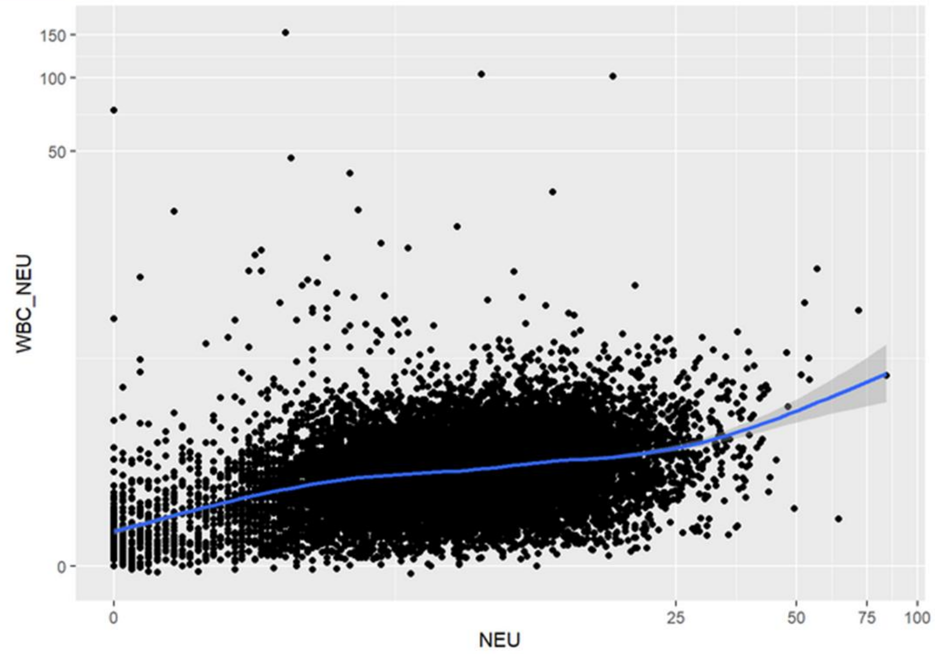
Possible consequences

Folks, I have an idea.

I‘ll redefine
WBC_noNEU = WBC – NEU
and use it with NEU in the model.

The correlation should vanish.
And I can still interpret the betas.

That was a good one!
(see also Gregorich, IJERPH 2021)



You will explain that to me?

Possible consequences

I wondered if those transformations affect my functional form selection…

Well they do, but perhaps you don‘t need to worry about functional forms…

Predictor	Selected functional form without pre-transformation	Selected functional form with log-pre-transformation
WBC_noNEU	Log	Sqrt
NEU	Sqrt	Identity
AGE	Identity	Identity
CREA	1/sqrt	Identity
PLT	Identity	Identity
BUN	Identity	Identity

Possible consequences

And that thing about missing values. There is multiple imputation...

Be careful! The relative importance of predictors will be different than with complete case analysis!

Possible consequences

Now with multiple imputation.
It does a good job for the complete predictors.

But not for CREA right?

Predictor	FDA	CCA	MIA
Intercept	0.386	0.554	0.388
WBC_noNEU	0.256	0.363	0.256
NEU	0.157	0.225	0.157
Age	0.208	0.297	0.210
CREA	**0.201**	**0.281**	**0.261**
PLT	0.032	0.046	0.032
BUN	0.185	0.264	0.221

$/\sqrt{2}$

$/\sqrt{1.16}$

Summary

IDA is the foundation for modeling: it complements domain expertise to support choice of model, its interpretation and presentation

IDA takes time and planning

- BUT: finding problems after modeling takes MORE time and may miss issues (not systematic)
- We provide worked example with code and workflow for this project

IDA needs to be reported in papers

- Transparency, rigor, reproducibility
- Suggestions in Huebner et al, BMC Med Res 2020
- Ten simple rules in Baillie et al, PLoS Comp Biol 2022

References

Remember!
IDA ≠ EDA

Comprehensive IDA Framework

Huebner M, le Cessie S, Schmidt CO, Vach W on behalf of STRATOS-TG3. A contemporary conceptual framework for initial data analysis. Observational Studies 2018; 4: 171-192. Link

Ten Simple Rules for IDA

Baillie et al on behalf of STRATOS TG3. Ten simple rules for Initial Data Annalysis. PLoS Comp Biol https://doi.org/10.1371/journal.pcbi.1009819

IDA Reporting

Huebner M, Vach W, le Cessie S, Schmidt C, Lusa L on behalf of STRATOS-TG3. Hidden Analyses: a review of reporting practice and recommendations for more transparent reporting of initial data analyses. BMC Med Res Meth 2020; 20:61. Link

An example

Lusa L, Huebner M. Organizing and Analyzing Data from the SHARE Study with an Application to Age and Sex Differences in Depressive Symptoms. IJERPH 2021;18(18):9684. doi: 10.3390/ijerph18189684.

Workflow: https://www.stratosida.org/

Data set: Ratzinger F, et al PlosOne 2014; Gregorich M et al, IJERPH 2021