Regression without regrets

M. Baillie, G. Heinze & M. Huebner

5/9/23

Table of contents

# Preface

The focus of this report is to provide guidance on conducting initial data analysis in a reproducible manner in the context of intended regression analyses.

# 1. IDA plan

This document exemplifies the pre-specified plan for initial data analysis (IDA plan) for the bacteremia study.

## 1.1 Prerequisites for the IDA plan

### 1.1.1 Analysis strategy

We assume that the aims of the study are to fit a diagnostic prediction model and to describe the functional form of each predictor. These aims are addressed by fitting a logistic regression model with bacteremia status as the dependent variable.

Based on domain expertise, the predictors are grouped by their assumed importance to predict bacteremia. Variables with known strong associations with bacteremia are age (AGE), leukocytes (WBC), blood urea neutrogen (BUN), creatinine (CREA), thrombocytes (PLT), and neutrophiles (NEU) and these predictors will be included in the model as key predictors. Predictors of medium importance are potassium (POTASS), and some acute-phase related parameters such as fibrinogen (FIB), C-reactive protein (CRP), aspartate transaminase (ASAT), alanine transaminase (ALAT), and gamma-glutamyl transpeptidase (GGT). All other predictors are of minor importance.

Continuous predictors should be modelled by allowing for flexible functional forms, where for all key predictors four degrees of freedom will be spent, and for predictors of medium and minor importance, three or two degrees of freedom should be foreseen at maximum, respectively. The decision on whether to use only key predictors, or to consider predictors also from the predictor sets of medium or minor importance depends on results of data screening, but will be made before uncovering the association of predictors with the outcome variable.

An adequate strategy to cope with missing values will also be chosen after screening the data. Candidate strategies are omission of predictors with abundant missing values, complete case analysis, single value imputation or multiple imputation with chained equations.

### 1.1.2 Analysis Ready Data dictionary

The **source** data dictionary of the bacteremia data set consists of columns for variable names, variable labels, scale of measurement (continuous or categorical), units, plausibility limits, and remarks. See [Appendix A](#sec-source-data).

An additional [**analysis ready** data dictionary](https://docs.google.com/spreadsheets/d/1Ft5eyenvDnMBoLvJmcBaklfrYcwyW-rkt-ivIkaphdA/edit#gid=1128598743) will be created to capture additional derivation, transformations and metadata relevant to the research question.

### 1.1.3 Domain expertise

The demographic variables age and sex are are chosen as the structural variables in this analysis for illustration purposes, since they are commonly considered important for describing a cohort in health studies. Key predictors and predictors of medium importance are as defined above. Laboratory analyses always bear the risk of machine failures, and hence missing values are a frequent challenge. This may differ between laboratory variables, but no a priori estimate about the expected proportion of missing values can be assumed. As most predictors measure concentrations of chemical compounds or cell counts, skewed distributions are expected. Some predictors describe related types of cells or chemical compounds, and hence some correlation between them is to be expected. For example, leukocytes consist of five different types of blood cells (BASO, EOS, NEU, LYM and MONO), and the sum of the concentration of these types approximately (but not exactly) gives the leukocyte count, which is recorded in the variable WBC. Moreover, these variables are given as absolute counts and as percentages of the sum of the five variables, which creates some correlation. Some laboratory variables differ by sex and age, but the special selection of patients for this study (suspicion of bacteremia) may distort or alter the expected correlations with sex and age.

## 1.2 IDA data derivations

Based on the prerequisites the following data derivations will be performed on the **source data** to form an **analysis ready data set** stored in the **data** folder.

The outcome variable from the source data **BC** will be renamed to **BACTEREMIA** to be more informative. A numeric variable **BACTEREMIAN** will also be derived to enable modelling with the following coding, 0: no, 1: yes.

For the purpose of stratifying IDA results by age, **AGE** will be categorized into the following three groups: [16, 50], (50, 65], (65, 101], both with numeric (**AGEGR01**) and character coding (**AGEGR01C**).

Predictors will be grouped by importance to the research question. Indicator flags for the predictor grouping will be derived:

* **KEY\_PRED\_FL01** will indicate the key predictors: AGE, WBC, BUN, CREA, PLT, and NEU
* **MED\_PRED\_FL01** will indicate the medium importance predictors: POTASS, FIB, CRP, ASAT, ALAT, GGT
* **REM\_PRED\_FL01** will indicator all remaining predictors not defined as key or medium importance.

An additional metadata flag **PARCAT01** will de derived to indicate the blood cell variables which form the Leukocytes predictor (**WBC**): BASO, EOS, NEU, LYM and MONO.

## 1.3 IDA planned analyses

The following sections detail the IDA planned analyses.

### 1.3.1 M1: Prevalence of missing values

Numbers and proportions of missing values will be reported for each predictor separately (M1). Type of missingness has not been recorded.

### 1.3.2 M2: Complete cases

The number of available complete cases (outcome and predictors) will be reported when considering:

1. the outcome variable (BACTEREMIA)
2. outcome and structural variables (BACTEREMIA, AGE, SEX)
3. outcome and key predictors only (BACTEREMIA, AGE, WBC, BUN, CREA, PLT, NEU)
4. outcome, key predictors and predictors of medium importance (BACTEREMIA, AGE, WBC, BUN, CREA, PLT, NEU, POTASS, FIB, CRP, ASAT, ALAT, GGT)
5. outcome and all predictors.

### 1.3.3 M3: Patterns of missing values

Patterns of missing values will be investigated by:

1. computing a table of complete cases (for the three predictor sets described above) for strata defined by the structural variables age and sex,
2. constructing a dendrogram of missing values to explore which predictors tend to be missing together.

### 1.3.4 U1: Univariate descriptions: categorical variables

For sex and bacteremia status, the frequency and proportion of each category will be described numerically.

### 1.3.5 U2: Univariate descriptions: continuous variables

For all continuous predictors, combo plots consisting of high-resolution histograms, boxplots and dotplots will be created. Because of the expected skew distribution, combo plots will also be created for log-transformed predictors.

As numerical summaries, minimum and maximum values, main quantiles (5th, 10th, 25th, 50th, 75th, 90th, 95th), and the first four moments (mean, standard deviation, skewness, curtosis) will be reported. The number of distinct values and the five most frequent values will be given, as well as the concentration ratio (ratio of frequency of most frequent value and mean frequency of each unique value).

Graphical and parametric multivariate analyses of the predictor space such as cluster analyses or the computation of variance inflation factors are heavily influenced by the distribution of the predictors. In order to make this set of analyses more robust to highly influential points or areas of the predictor support, some predictors may need transformation (e.g. logarithmic). We will compute the correlation of the untransformed and log-transformed predictors with normal deviates. Since some predictors may have values at or close to 0, we will consider the pseudolog transformation (Johnson 1949) which provides a smooth transition from linear (close to 0) to logarithmic (further away from 0). The transformation has a parameter which we will optimize separately for each predictor in order to achieve an optimal approximation to a normal distribution monitored via the correlation of normal deviates with the transformed predictor. For those predictors for which the pseudolog-transformation increases correlation with normal deviates by at least 0.2 units of the correlation coefficient, the pseudolog-transformed predictor will be used in multivariate IDA instead of the respective original predictor. For those predictors, histograms and boxplots will be provided on both the original and the transformed scale.

### 1.3.6 V1: Multivariate descriptions: associations of predictors with structural variables

A scatterplot of each predictor with age, with different panels for males and females will be constructed. Associated Spearman correlation coefficients will be computed.

### 1.3.7 V2: Multivariate descriptions: correlation analyses

A matrix of Spearman correlation coefficients between all pairs of predictors will be computed and described numerically as well as by means of a heatmap.

### 1.3.8 VE1: Multivariate descriptions: comparing nonparametric and parametric predictor correlation

A matrix of Pearson correlation coefficients will be computed. Predictor pairs for which Spearman and Pearson correlation coefficients differ by more than 0.1 correlation units will be depicted in scatterplots.

### 1.3.9 VE2: Variable clustering

A variable clustering analysis will be performed to evaluate which predictors are closely associated. A dendrogram groups predictors by their correlation. Scatterplots of pairs of predictors with Spearman correlation coefficients greater than 0.8 will be created.

### 1.3.10 VE3: Redundancy

Variance inflation factors will be computed between the candidate predictors. This will be done for the three possible candidate models, and using all complete cases in the respective candidate predictor sets. Redundancy will further be explored by computing parametric additive models for each predictor in the three candidate models.

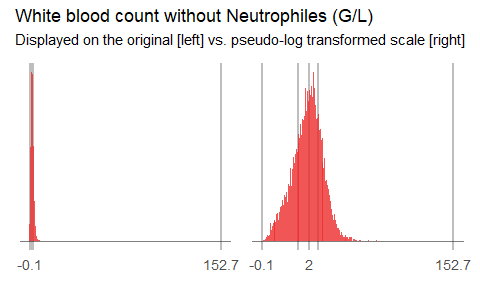
# 2. Final Analysis ready data

## 2.1 Final dervivations and transformations

From the IDA exercise a decision was made to derive a new variable based on White Blood Count in the absence of Neutrophiles:

* WBC\_NEU: WBC - NEU

The new derivation will also be checked as a candidate for pseudo-log transformation.



The new derivation and transformation is added to the analysis ready data, setting the correct indicator flags to enable final analyses and modelling.

Check flags

[1] "c(\"PLT\", \"CREA\", \"BUN\", \"NEU\", \"AGE\", \"WBC\_noNEU\")"

## 2.2 Save final analysis ready data sets

The final analysis ready data sets are then stored in data/.

# References

Johnson, N. L. 1949. “Systems of Frequency Curves Generated by Methods of Translation.” *Biometrika* 36 (1/2): 149–76. <http://www.jstor.org/stable/2332539>.

# Appendix A — Source data

The **source** data available in this repository (**DC 2019-0054**) can be located in the data-raw folder. For persistance, the data set is also published on [Zenodo](https://doi.org/10.5281/zenodo.7554815) with the following doi: https://doi.org/10.5281/zenodo.7554815.

The following sections provide a short overview of the data dictionary which accompanies the source data, and a short description of the data.

## A.1 Data dictionary

The data dictionary provides an overview of the collected data. First, we read and display the data dictionary below providing an overview of the collected measurements.

The variable name and label are displayed alongside the measurement scale and units as well as remarks and relevant study information from\_paper.

| variable | label | scale\_of\_measurement | units |
| --- | --- | --- | --- |
| ID | Patient Identification | nominal | 1-14691 |
| SEX | Patient sex | nominal | 1=male, 2=female |
| AGE | Patient Age | continuous | years |
| MCV | Mean corpuscular volume | continuous | pg |
| HGB | Haemoglobin | continuous | G/L |
| HCT | Haematocrit | continuous | % |
| PLT | Blood platelets | continuous | G/L |
| MCH | Mean corpuscular hemoglobin | continuous | fl |
| MCHC | Mean corpuscular hemoglobin concentration | continuous | g/dl |
| RDW | Red blood cell distribution width | continuous | % |
| MPV | Mean platelet volume | continuous | fl |
| LYM | Lymphocytes | continuous | G/L |
| MONO | Monocytes | continuous | G/L |
| EOS | Eosinophils | continuous | G/L |
| BASO | Basophiles | continuous | G/L |
| NT | Normotest | continuous | % |
| APTT | Activated partial thromboplastin time | continuous | sec |
| FIB | Fibrinogen | continuous | mg/dl |
| SODIUM | Sodium | continuous | mmol/L |
| POTASS | Potassium | continuous | mmol/L |
| CA | Calcium | continuous | mmol/L |
| PHOS | Phosphate | continuous | mmol/L |
| MG | Magnesium | continuous | mmol/L |
| CREA | Creatinine | continuous | mg/dl |
| BUN | Blood urea nitrogen | continuous | mg/dl |
| HS | Uric acid | continuous | mg/dl |
| GBIL | Bilirubin | continuous | mg/dl |
| TP | Total protein | continuous | G/L |
| ALB | Albumin | continuous | G/L |
| AMY | Amylase | continuous | U/L |
| PAMY | Pancreas amylase | continuous | U/L |
| LIP | Lipases | continuous | U/L |
| CHE | Cholinesterase | continuous | kU/L |
| AP | Alkaline phosphatase | continuous | U/L |
| ASAT | Aspartate transaminase | continuous | U/L |
| ALAT | Alanin transaminase | continuous | U/L |
| GGT | Gamma-glutamyl transpeptidase | continuous | G/L |
| LDH | Lactate dehydrogenase | continuous | U/L |
| CK | Creatinine kinases | continuous | U/L |
| GLU | Glucoses | continuous | mg/dl |
| TRIG | Triclyceride | continuous | mg/dl |
| CHOL | Cholesterol | continuous | mg/dl |
| CRP | C-reactive protein | continuous | mg/dl |
| BASOR | Basophile ratio | continuous | % |
| EOSR | Eosinophil ratio | continuous | % |
| LYMR | Lymphocyte ratio | continuous | % (mg/dl) |
| MONOR | Monocyte ratio | continuous | % |
| NEU | Neutrophiles | continuous | G/L |
| NEUR | Neutrophile ratio | continuous | % |
| PDW | Platelet distribution width | continuous | % |
| RBC | Red blood count | continuous | T/L |
| WBC | White blood count | continuous | G/L |
| BloodCulture | Blood culture result for bacteremia | nominal | no, yes |

## A.2 Source data

We also display a short snapshot of source data set from the data-raw folder of the project directory. The snapshot provides a glimpse of the data, giving the data dictionary more context.

We do not display all observations measured as it is too wide and long to fit reasonably in to the report. However, we refer you to the [Zenodo page](https://doi.org/10.5281/zenodo.7554815) for an interactive overview of the source data.

Rows: 14,691  
Columns: 53  
$ ID <dbl> 1, 3, 5, 7, 9, 10, 11, 12, 13, 19, 21, 22, 23, 25, 26, 27~  
$ SEX <dbl> 2, 1, 1, 1, 2, 1, 1, 1, 1, 2, 1, 1, 2, 2, 2, 1, 2, 2, 2, ~  
$ AGE <dbl> 62, 72, 46, 84, 38, 68, 55, 55, 67, 52, 47, 29, 59, 51, 4~  
$ MCV <dbl> 99.3, 85.1, 96.3, 91.3, 85.1, 104.5, 99.3, 77.0, 95.3, 83~  
$ HGB <dbl> 11.5, 10.3, 7.4, 10.3, 13.7, 15.7, 14.6, 10.8, 10.9, 10.3~  
$ HCT <dbl> 35.9, 34.7, 22.8, 31.1, 38.7, 46.9, 43.5, 34.8, 30.4, 30.~  
$ PLT <dbl> 307, 182, 64, 309, 183, 144, 242, 38, 88, 105, 216, 188, ~  
$ MCH <dbl> 31.5, 26.0, 31.2, 30.4, 30.2, 34.8, 33.1, 23.8, 33.6, 28.~  
$ MCHC <dbl> 31.8, 30.6, 32.4, 33.3, 35.3, 33.5, 33.4, 30.5, 35.3, 34.~  
$ RDW <dbl> 19.5, 15.0, 19.7, 13.8, 12.6, 13.9, 13.1, 16.8, 13.3, 13.~  
$ MPV <dbl> 10.8, 9.7, 11.1, 8.5, 10.0, 10.9, 10.3, NA, 10.7, 11.3, 1~  
$ LYM <dbl> 0.4, 0.4, 1.5, 1.3, 0.8, 2.2, 2.1, 0.4, 0.4, 0.9, 0.7, 1.~  
$ MONO <dbl> 1.7, 0.2, 1.2, 0.8, 0.4, 0.9, 1.6, 0.1, 0.2, 0.9, 0.6, 0.~  
$ EOS <dbl> 0.0, 0.1, 0.1, 0.0, 0.0, 0.1, 0.3, 0.1, 0.0, 0.3, 0.0, 0.~  
$ BASO <dbl> 0.1, 0.0, 0.1, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.1, 0.1, 0.~  
$ NT <dbl> 86, 90, 58, 67, 95, 61, NA, 93, 57, 69, 108, 86, 93, 83, ~  
$ APTT <dbl> 28.8, 29.8, 36.3, 38.2, 33.1, 41.8, NA, 36.3, 33.8, 28.1,~  
$ FIB <dbl> 578, NA, 313, 487, 490, 400, NA, 413, 431, 407, 604, 476,~  
$ SODIUM <dbl> 137, 141, 147, 141, 137, 141, 139, 142, 143, 136, 131, 13~  
$ POTASS <dbl> 3.88, NA, 4.61, 4.71, NA, 4.41, 3.69, 4.67, 2.35, 3.80, 5~  
$ CA <dbl> 2.29, 2.21, 1.92, 2.05, 2.34, 2.08, NA, 2.31, 2.10, 1.92,~  
$ PHOS <dbl> 1.20, 0.58, 1.51, 2.17, 0.97, 0.99, NA, 1.16, 0.51, 0.72,~  
$ MG <dbl> 0.66, NA, 1.03, 0.83, 0.74, 0.56, NA, 0.87, 0.36, 0.53, 0~  
$ CREA <dbl> 0.65, 0.76, 1.25, 2.78, 0.65, 0.82, 1.21, 1.77, 1.00, 0.5~  
$ BUN <dbl> 5.7, 19.9, 50.6, 47.5, 8.5, 15.3, 13.0, 29.8, 15.0, 14.0,~  
$ HS <dbl> 5.3, NA, NA, 9.7, 3.0, 5.5, NA, 6.2, 4.7, 4.0, 4.0, 4.1, ~  
$ GBIL <dbl> 0.59, 0.48, 8.42, 0.35, 0.42, 2.40, 1.13, 0.45, 1.21, 2.4~  
$ TP <dbl> 67.0, 65.3, 40.5, 61.2, 78.4, 57.5, NA, 70.8, 67.4, 53.8,~  
$ ALB <dbl> 36.7, 37.4, 22.1, 33.2, 43.8, 30.1, NA, 43.6, 35.4, 24.8,~  
$ AMY <dbl> 30, NA, 146, 92, 84, 95, 117, 177, NA, 35, 79, 16, 25, 32~  
$ PAMY <dbl> 16, NA, NA, 28, 50, 57, NA, 43, NA, 35, 63, 14, 15, 20, 3~  
$ LIP <dbl> 10, NA, 89, 18, 50, 25, 73, 30, NA, 38, 52, 19, 14, 26, 5~  
$ CHE <dbl> 5.12, 5.61, 2.52, 4.10, 6.91, 6.79, NA, 7.40, NA, 2.64, 2~  
$ AP <dbl> 85, 80, 119, 94, 108, 68, 51, 153, 239, 146, 180, 64, 74,~  
$ ASAT <dbl> 22, 28, 124, 774, 35, 32, 29, 26, 91, 97, 24, 13, 25, 31,~  
$ ALAT <dbl> 14, 25, 135, 72, 22, 11, 20, 32, 57, 156, 63, 23, 27, 53,~  
$ GGT <dbl> 48, 61, 134, 23, 72, 68, 138, 96, 446, 192, 266, 19, 66, ~  
$ LDH <dbl> 284, NA, 696, 1787, NA, 263, 303, 181, 183, 277, 221, 299~  
$ CK <dbl> 23, 36, 40, 2422, 79, 75, 230, 87, 53, 87, 30, 118, 17, 1~  
$ GLU <dbl> 107, 84, 107, 105, 93, 89, 91, 96, 86, 104, 104, 102, 161~  
$ TRIG <dbl> 105, NA, NA, 134, 152, 85, NA, 129, 62, 207, 292, 221, 12~  
$ CHOL <dbl> 175, NA, NA, 141, 167, 144, NA, 156, 118, 123, 194, 151, ~  
$ CRP <dbl> 3.94, 1.42, 12.09, 3.78, 11.17, 5.89, 17.84, 1.29, 1.36, ~  
$ BASOR <dbl> 0.4132231, 0.0000000, 0.5681818, 0.0000000, 0.0000000, 0.~  
$ EOSR <dbl> 0.0000000, 0.8264463, 0.5681818, 0.0000000, 0.0000000, 1.~  
$ LYMR <dbl> 1.652893, 3.305785, 8.522727, 11.016949, 8.333333, 22.000~  
$ MONOR <dbl> 7.024793, 1.652893, 6.818182, 6.779661, 4.166667, 9.00000~  
$ NEU <dbl> 22.0, 11.4, 14.7, 9.7, 8.4, 6.8, 8.9, 1.2, NA, 3.8, 8.2, ~  
$ NEUR <dbl> 90.90909, 94.21488, 83.52273, 82.20339, 87.50000, 68.0000~  
$ PDW <dbl> 10.6, 11.4, 14.1, 8.7, 12.2, 12.9, 12.5, NA, NA, 13.2, 12~  
$ RBC <dbl> 3.7, 3.9, 2.5, 3.5, 4.4, 4.3, 4.5, 4.7, NA, 3.5, 3.3, 2.5~  
$ WBC <dbl> 24.10, 12.17, 17.45, 11.58, 9.86, 9.94, 13.06, 1.78, NA, ~  
$ BloodCulture <chr> "no", "no", "no", "no", "no", "no", "no", "no", "yes", "n~