Regression without regrets

M. Baillie, G. Heinze & M. Huebner

3/22/23

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. Bacteremia study

## 1.1 Bacteremia study overview

We will exemplify our proposed systematic approach to data screening by means of a diagnostic study with the primary aim of using age, sex and 49 laboratory variables to fit a diagnostic prediction model for the bacteremia status (= presence of bacteria in the blood stream) of a blood sample. A secondary aim of the study is to describe the functional form of each predictor in the model. Between January 2006 and December 2010, patients with the clinical suspicion to suffer from bacteremia were included if blood culture analysis was requested by the responsible physician and blood was sampled for assessment of hematology and biochemistry. An analysis of this study can be found in Ratzinger et al. (2014).

The data consists of 14,691 observations from different patients and 51 potential predictors. To protect data privacy our version of this data was slightly modified compared to the original version, and this modified version was cleared by the Medical University of Vienna for public use (**DC 2019-0054**). Compared to the official results given in (Ratzinger et al. 2014), our results may differ to a negligible degree.

## 1.2 Source dataset

### 1.2.1 Where to access the data?

We refer to the **source** data as the raw data set available in this repository (**DC 2019-0054**). The data set is published on [Zenodo](https://doi.org/10.5281/zenodo.7554815) with the following doi: https://doi.org/10.5281/zenodo.7554815.

For simplicity, we have also stored the *source* data and accompanying materials such as the **data dictionary** the data-raw directory.

### 1.2.2 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 | remark | from\_paper |
| --- | --- | --- | --- | --- | --- |
| ID | Patient Identification | nominal | 1-14691 | NA | NA |
| SEX | Patient sex | nominal | 1=male, 2=female | NA | Female: Male |
| AGE | Patient Age | continuous | years | Alter=German age | Albumin (G/L) 14,187 33.7 (28?39.3) 32 (26.925?36.7) ,0.0001 0.568 |
| MCV | Mean corpuscular volume | continuous | pg | NA | MCV (pg) 15,941 88.1 (84.6?91.9) 88.6 (84.8?92.5) 0.0044 0.524 |
| HGB | Haemoglobin | continuous | G/L | NA | Haemoglobin(G/L) 15,942 11.4 (9.9?13.2) 11.1 (9.5?12.6) ,0.0001 0.554 |
| HCT | Haematocrit | continuous | % | NA | Haematocrit (%) 15,941 34.4 (29.8?39.2) 33.1 (28.5?37.5) ,0.0001 0.561 |
| PLT | Blood platelets | continuous | G/L | NA | PLT (G/L) 15,940 206 (142?279.25) 180.5 (115?248) ,0.0001 0.575 |
| MCH | Mean corpuscular hemoglobin | continuous | fl | NA | MCH (fl) 15,941 29.7 (28.3?30.9) 29.8 (28.5?31.2) 0.0019 0.526 |
| MCHC | Mean corpuscular hemoglobin concentration | continuous | g/dl | NA | MCHC (g/dl) 15,941 33.5 (32.6?34.4) 33.6 (32.7?34.5) n.s. |
| RDW | Red blood cell distribution width | continuous | % | NA | RDW (%) 15,924 14.4 (13.3?15.925) 14.9 (13.7?16.6) ,0.0001 0.572 |
| MPV | Mean platelet volume | continuous | fl | NA | MPV (fl) 15,214 10.3 (9.7?11) 10.4 (9.7?11.1) n.s. |
| LYM | Lymphocytes | continuous | G/L | NA | Lymphocytes (G/L) 15,695 1.1 (0.7?1.6) 0.7 (0.4?1.1) ,0.0001 0.683 |
| MONO | Monocytes | continuous | G/L | NA | Monocytes (G/L) 15,710 0.8 (0.5?1.1) 0.6 (0.3?1) ,0.0001 0.598 |
| EOS | Eosinophils | continuous | G/L | NA | Eosinophils (G/L) 15,373 0.6 (0.1?1.8) 0.2 (0?0.8) ,0.0001 0.641 |
| BASO | Basophiles | continuous | G/L | NA | Basophiles % 15,375 0.2 (0.1?0.3) 0.1 (0.1?0.2) ,0.0001 0.606 |
| NT | Normotest | continuous | % | Measures thromboplastin time | Normotest (%) 13,339 84 (67?101) 78 (60?94) ,0.0001 0.571 |
| APTT | Activated partial thromboplastin time | continuous | sec | NA | aPTT (sec) 13,251 37.8 (34.2?42.8) 37.8 (34.2?43) n.s. |
| FIB | Fibrinogen | continuous | mg/dl | NA | Fibrinogen (mg/dl) 13,211 526 (393?667) 546 (424?701) 0.0001 0.538 |
| SODIUM | Sodium | continuous | mmol/L | Natrium=German sodium | Sodium (mmol/L) 14,542 138 (135?140) 136 (133?139) ,0.0001 0.602 |
| POTASS | Potassium | continuous | mmol/L | NA | Potassium (mmol/L) 13,774 3.95 (3.67?4.3) 3.97 (3.595?4.365) n.s. |
| CA | Calcium | continuous | mmol/L | NA | Calcium (mmol/L) 14,592 2.23 (2.09?2.35) 2.21 (2.08?2.33) 0.0001 0.533 |
| PHOS | Phosphate | continuous | mmol/L | NA | Phosphate(mmol/L) 14,664 1 (0.81?1.2) 0.95 (0.76?1.19) ,0.0001 0.537 |
| MG | Magnesium | continuous | mmol/L | NA | MG (mmol/L) 13,989 0.81 (0.73?0.89) 0.77 (0.68?0.86) ,0.0001 0.582 |
| CREA | Creatinine | continuous | mg/dl | NA | Creatinine (mg/dl) 15,813 0.99 (0.81?1.31) 1.2 (0.89?1.87) ,0.0001 0.611 |
| BUN | Blood urea nitrogen | continuous | mg/dl | NA | BUN (mg/dl) 15,800 16.2 (11.4?25.8) 22.5 (14.7?37.78) ,0.0001 0.633 |
| HS | Uric acid | continuous | mg/dl | Harns?ure=German Uric acid | Uric acid (mg/dl) 12,709 5 (3.7?6.5) 5.5 (3.9?7.6) ,0.0001 0.562 |
| GBIL | Bilirubin | continuous | mg/dl | NA | Bilirubin (mg/dl) 14,431 0.75 (0.52?1.19) 1.02 (0.66?1.73) ,0.0001 0.621 |
| TP | Total protein | continuous | G/L | NA | TP (G/L) 14,301 65.8 (56.8?73.4) 64.7 (56.4?71.5) 0.0019 0.528 |
| ALB | Albumin | continuous | G/L | NA | ALAT (U/L) 14,919 26 (16?47) 30 (18?60) ,0.0001 0.55 |
| AMY | Amylase | continuous | U/L | NA | Amylase (U/L) 11,783 50 (34?77) 44 (28?70) ,0.0001 0.565 |
| PAMY | Pancreas amylase | continuous | U/L | NA | PAMY (U/L) |
| LIP | Lipases | continuous | U/L | NA | Lipases (U/L) 11,988 23 (13?40) 22 (12?38) n.s. |
| CHE | Cholinesterase | continuous | kU/L | NA | CHE (kU/L) 13,353 4.66 (3.2?6.29) 3.94 (2.66?5.48) ,0.0001 0.591 |
| AP | Alkaline phosphatase | continuous | U/L | NA | ALP (U/L) 14,479 83 (62?120) 100 (72?164) ,0.0001 0.601 |
| ASAT | Aspartate transaminase | continuous | U/L | NA | ASAT (U/L) 14,745 31 (22?56) 37 (24?70.25) ,0.0001 0.558 |
| ALAT | Alanin transaminase | continuous | U/L | NA | Age 15,985 58 (42?69) 65 (53?74) ,0.0001 0.611 |
| GGT | Gamma-glutamyl transpeptidase | continuous | G/L | NA | GGT (G/L) 14,629 48 (25?112) 73 (35?180) ,0.0001 0.599 |
| LDH | Lactate dehydrogenase | continuous | U/L | NA | LDH (U/L) 14,150 239 (186?334) 249 (199?331.5) 0.0037 0.527 |
| CK | Creatinine kinases | continuous | U/L | NA | CK (U/L) 13,763 82 (42?190) 67 (34?142) ,0.0001 0.557 |
| GLU | Glucoses | continuous | mg/dl | NA | Glucoses (mg/dl) 11,350 113 (96?137) 121 (99?154) ,0.0001 0.559 |
| TRIG | Triclyceride | continuous | mg/dl | NA | Triglyceride (mg/dl) 10,549 115 (83?164) 118 (85?170) n.s. |
| CHOL | Cholesterol | continuous | mg/dl | NA | Cholesterol (mg/dl) 10,565 146 (114?183) 132 (105?171) ,0.0001 0.564 |
| CRP | C-reactive protein | continuous | mg/dl | NA | CRP (mg/dl) 15,820 8.39 (2.77?16.15) 11.68 (5.22?21.19) ,0.0001 0.596 |
| BASOR | Basophile ratio | continuous | % | NA | Basophiles (G/L) 15,827 0 (0?0) 0 (0?0) ,0.0001 0.47 |
| EOSR | Eosinophil ratio | continuous | % | NA | Eosinophil % 15,831 0.1 (0?0.2) 0 (0?0.1) ,0.0001 0.626 |
| LYMR | Lymphocyte ratio | continuous | % (mg/dl) | NA | Lymphocytes % (mg/dl) 15,250 11.6 (7.1?18.6) 7 (4.15?12.2) ,0.0001 0.674 |
| MONOR | Monocyte ratio | continuous | % | NA | Monocytes % 15,268 8.1 (5.8?10.7) 6.1 (3.5?8.8) ,0.0001 0.645 |
| NEU | Neutrophiles | continuous | G/L | NA | Neutrophiles (G/L) 15,181 7.3 (4.6?10.7) 8.4 (5.23?12.7) ,0.0001 0.559 |
| NEUR | Neutrophile ratio | continuous | % | NA | Neutrophiles % 15,181 77.7 (68.7?84.6) 85.8 (78.3?90.5) ,0.0001 0.696 |
| PDW | Platelet distribution width | continuous | % | NA | PDW (%) 14,776 12 (10.8?13.4) 12.1 (10.8?13.7) n.s. |
| RBC | Red blood count | continuous | T/L | NA | RBC (T/L) 15,478 3.9 (3.4?4.5) 3.7 (3.2?4.2) ,0.0001 0.567 |
| WBC | White blood count | continuous | G/L | NA | WBC (G/L) 15,477 9.58 (6.64?13.46) 10.205 (6.61?14.86) n.s. |
| BloodCulture | Blood culture result for bacteremia | nominal | no, yes | NA | NA |

### 1.2.3 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~

Ratzinger, Franz, Michel Dedeyan, Matthias Rammerstorfer, Thomas Perkmann, Heinz Burgmann, Athanasios Makristathis, Georg Dorffner, Felix Lötsch, Alexander Blacky, and Michael Ramharter. 2014. “A Risk Prediction Model for Screening Bacteremic Patients: A Cross Sectional Study.” *PLOS ONE* 9 (9): 1–10. <https://doi.org/10.1371/journal.pone.0106765>.