Regression without regrets

M. Baillie, G. Heinze & M. Huebner

3/7/23

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. Univariate distribution checks

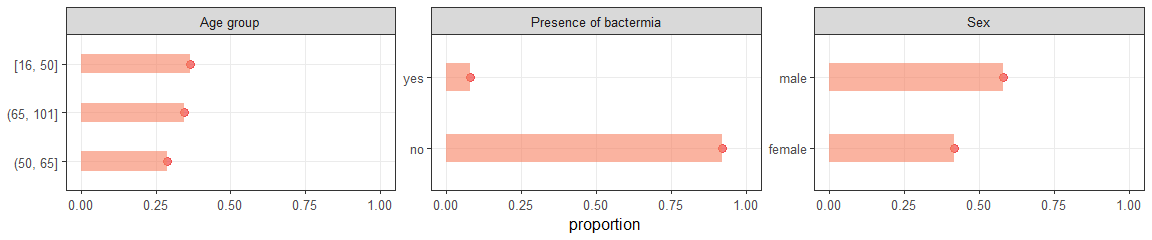
This section reports a series of univariate summary checks of the bacteremia dataset.

## 1.1 U1: Categorical variables

Age group, sex and bactermia status are described by frequencies and proportions in each category.

| Category | Count | Proportion |
| --- | --- | --- |
| Age group |
| (50, 65] | 4250 | 0.29 |
| (65, 101] | 5076 | 0.35 |
| [16, 50] | 5365 | 0.37 |
| Sex |
| female | 6155 | 0.42 |
| male | 8536 | 0.58 |
| Presence of bactermia |
| no | 13511 | 0.92 |
| yes | 1180 | 0.08 |

Also plot the categories.



## 1.2 Continuous variables

### 1.2.1 U2: Univariate distributions of continuous variables

#### 1.2.1.1 U2: Structural variables

The only structural continuous variables is AGE. This variable is also a key predictor (see below).

#### 1.2.1.2 U2: Key predictors

Note: the structural variable Age is also considered a key predictor.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

|  |  |
| --- | --- |
|  |  |

#### 1.2.1.3 U2: Predictors of medium importance

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

|  |  |
| --- | --- |
|  |  |

#### 1.2.1.4 U2: Remaining predictors

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |

|  |  |
| --- | --- |
|  |  |

### 1.2.2 Numerical summaries

#### 1.2.2.1 Key predictors

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14649 | 42 | 718 | 1 | 220 | 130.1 | 50 | 81 | 140 | 204 | 277 | 369 | 445 |

lowest : 0 1 2 3 4 , highest: 1068 1211 1321 1639 2092

CREA: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,iVBORw0KGgoAAAANSUhEUgAAAJcAAAANCAMAAACTvAxuAAAACVBMVEUAAADMzMz////1iUV5AAAAWUlEQVQ4je2TMQoAIAwDE///aG3sIAqCWwqeQ6mKHoWAAqAXaAKD5sT3euN7vZE5DC+rSM4pSctqYBW8nMxqePm4RQjB1Usd93LuXK5pUU9nZX6TzT36cdoBj3gMym+2tjYAAAAASUVORK5CYII=)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 14532 | 159 | 674 | 1 | 1.329 | 0.8518 | 0.620 | 0.690 | 0.810 | 1.000 | 1.350 | 2.160 | 3.144 |

lowest : 0.26 0.27 0.28 0.29 0.30 , highest: 15.24 15.40 15.67 16.64 20.75

BUN: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,iVBORw0KGgoAAAANSUhEUgAAAJcAAAANCAMAAACTvAxuAAAACVBMVEUAAADMzMz////1iUV5AAAAZElEQVQ4jdXTwQrAIAwD0NT//+jhaiZ0U7wlCz1IT49iEAx6wiVozO1qLvmJywZWXS4yMgAzF///jLSHzPteHidLA2rEqqVLTlu7tDbsWEJhb98BCzFn//xalW0WbsxT/1yMV1y21QwKyo8xNQAAAABJRU5ErkJggg==)

| 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

NEU: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,iVBORw0KGgoAAAANSUhEUgAAAJcAAAANCAMAAACTvAxuAAAACVBMVEUAAADMzMz////1iUV5AAAAZ0lEQVQ4jc2VSw7AIAgFx97/0A19JUVru+WNnwCriRFlJKDlAUcCWh5Ur8BFbfXC18tCLRXAzCvv/0RrK8rn1nvTfF6+Xnq4dnR7/eHqRc9by9qKG+LbvMaz1bgWRkzFJZ077TOpxRPORQvfG+yn6gAAAABJRU5ErkJggg==)

| n | missing | distinct | Info | Mean | Gmd | .05 | .10 | .25 | .50 | .75 | .90 | .95 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 13963 | 728 | 374 | 1 | 8.367 | 5.776 | 1.60 | 2.70 | 4.60 | 7.30 | 10.80 | 15.08 | 18.40 |

lowest : 0.0 0.1 0.2 0.3 0.4 , highest: 54.0 56.4 63.7 71.6 83.8

WBC: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,iVBORw0KGgoAAAANSUhEUgAAAIIAAAANCAMAAABSO0bCAAAACVBMVEUAAADMzMz////1iUV5AAAAVklEQVQ4je2TSwoAIAhEx+5/6PwRSdB2XDiRVG4e0oNoACEGSwOvrDRBwCA0QHAhqErkFJhjGIQbgciAJGAihBBPQhNb4hf7uVLM8d45Rwm/6vPfBt0b3GoLNRPwTDwAAAAASUVORK5CYII=)

| 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: Parameter analysis value (Numeric)

![image](data:image/png;base64;base64,iVBORw0KGgoAAAANSUhEUgAAAJcAAAANCAMAAACTvAxuAAAACVBMVEUAAADMzMz////1iUV5AAAAeUlEQVQ4jcXTSw4AERAE0Gr3P/REhzDM+PSHWohVe0iB7IMczYxgGSCgimaSpemF0tFUrnxuz1HLpK4hpZWlxcNVTeXtjkvyevPOgHsldXyEVoraXCDdqby9X9b/0VMhd51VzWi4hhrLLrt+cdw1y7YJEgtalughegBpCgbX7iPW9QAAAABJRU5ErkJggg==)

| 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 |

#### 1.2.2.2 Predictors of medium importance

### 1.2.3 Data Frame Summary

#### 1.2.3.1 dat

**Dimensions:** 14691 x 6  
**Duplicates:** 40

| Variable | Label | Stats / Values | Freqs (% of Valid) | Graph | Missing |
| --- | --- | --- | --- | --- | --- |
| FIB [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 547.36 (208.13) min < med < max: 55 < 529 < 1593 IQR (CV) : 277 (0.38) | 1084 distinct values |  | 2567 (17.5%) |
| POTASS [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 4 (0.64) min < med < max: 1.92 < 3.95 < 36.62 IQR (CV) : 0.63 (0.16) | 408 distinct values |  | 2008 (13.7%) |
| ASAT [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 86.9 (404.69) min < med < max: 3 < 31 < 13991 IQR (CV) : 34 (4.66) | 650 distinct values |  | 1154 (7.9%) |
| ALAT [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 67.66 (311.05) min < med < max: 0 < 26 < 15059 IQR (CV) : 32 (4.6) | 578 distinct values |  | 987 (6.7%) |
| GGT [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 115.06 (208.98) min < med < max: 3 < 49 < 5171 IQR (CV) : 92 (1.82) | 858 distinct values |  | 1262 (8.6%) |
| CRP [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 10.92 (9.58) min < med < max: 0 < 8.57 < 76.32 IQR (CV) : 13.58 (0.88) | 3328 distinct values |  | 155 (1.1%) |

#### 1.2.3.2 Remaining predictors

### 1.2.4 Data Frame Summary

#### 1.2.4.1 dat

**Dimensions:** 14691 x 38  
**Duplicates:** 0

| Variable | Label | Stats / Values | Freqs (% of Valid) | Graph | Missing |
| --- | --- | --- | --- | --- | --- |
| MCV [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 88.35 (6.46) min < med < max: 51 < 88.3 < 128.7 IQR (CV) : 7.3 (0.07) | 506 distinct values |  | 42 (0.3%) |
| HGB [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 11.57 (2.25) min < med < max: 3 < 11.4 < 21 IQR (CV) : 3.3 (0.19) | 157 distinct values |  | 41 (0.3%) |
| HCT [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 34.48 (6.51) min < med < max: 0 < 34.3 < 66.6 IQR (CV) : 9.3 (0.19) | 404 distinct values |  | 42 (0.3%) |
| MCH [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 29.58 (2.53) min < med < max: 14.9 < 29.7 < 47.4 IQR (CV) : 2.6 (0.09) | 232 distinct values |  | 42 (0.3%) |
| MCHC [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 33.47 (1.4) min < med < max: 23.7 < 33.5 < 43.5 IQR (CV) : 1.8 (0.04) | 124 distinct values |  | 42 (0.3%) |
| RDW [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 15 (2.29) min < med < max: 10.6 < 14.5 < 31.8 IQR (CV) : 2.6 (0.15) | 173 distinct values |  | 56 (0.4%) |
| MPV [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 10.38 (1.01) min < med < max: 7.3 < 10.3 < 15 IQR (CV) : 1.3 (0.1) | 71 distinct values |  | 702 (4.8%) |
| LYM [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 1.37 (7.46) min < med < max: 0 < 1 < 578.1 IQR (CV) : 0.9 (5.46) | 114 distinct values |  | 262 (1.8%) |
| MONO [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 0.85 (0.65) min < med < max: 0 < 0.8 < 20.4 IQR (CV) : 0.6 (0.76) | 67 distinct values |  | 246 (1.7%) |
| EOS [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 0.11 (0.27) min < med < max: 0 < 0.1 < 15.8 IQR (CV) : 0.1 (2.34) | 36 distinct values |  | 135 (0.9%) |
| BASO [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 0.02 (0.08) min < med < max: 0 < 0 < 6.5 IQR (CV) : 0 (4.71) | 18 distinct values |  | 146 (1.0%) |
| NT [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 83.22 (27.15) min < med < max: 4 < 83 < 152 IQR (CV) : 34 (0.33) | 149 distinct values |  | 2467 (16.8%) |
| APTT [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 40.06 (10.97) min < med < max: 21.4 < 37.7 < 176.1 IQR (CV) : 8.6 (0.27) | 631 distinct values |  | 2549 (17.4%) |
| SODIUM [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 137.21 (4.73) min < med < max: 106 < 137 < 170 IQR (CV) : 5 (0.03) | 58 distinct values |  | 1282 (8.7%) |
| CA [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 2.21 (0.2) min < med < max: 1.03 < 2.22 < 4.4 IQR (CV) : 0.26 (0.09) | 185 distinct values |  | 1276 (8.7%) |
| PHOS [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 1.05 (0.4) min < med < max: 0.3 < 0.99 < 6.22 IQR (CV) : 0.39 (0.38) | 306 distinct values |  | 1242 (8.5%) |
| MG [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 0.81 (0.15) min < med < max: 0.2 < 0.81 < 2.22 IQR (CV) : 0.17 (0.19) | 146 distinct values |  | 1869 (12.7%) |
| HS [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 5.41 (2.45) min < med < max: 1.3 < 5 < 22.7 IQR (CV) : 2.9 (0.45) | 169 distinct values |  | 3061 (20.8%) |
| GBIL [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 1.41 (2.75) min < med < max: 0.11 < 0.77 < 51.77 IQR (CV) : 0.7 (1.96) | 885 distinct values |  | 1441 (9.8%) |
| TP [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 64.9 (11.46) min < med < max: 29.9 < 65.7 < 120.9 IQR (CV) : 16.4 (0.18) | 649 distinct values |  | 1583 (10.8%) |
| ALB [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 33.42 (7.44) min < med < max: 10 < 33.6 < 55.7 IQR (CV) : 11.2 (0.22) | 401 distinct values |  | 1676 (11.4%) |
| AMY [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 90.83 (805.23) min < med < max: 8 < 49 < 56146 IQR (CV) : 43 (8.87) | 488 distinct values |  | 3913 (26.6%) |
| PAMY [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 41.66 (448.01) min < med < max: 1 < 22 < 38369 IQR (CV) : 22 (10.75) | 280 distinct values |  | 7114 (48.4%) |
| LIP [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 63.82 (603.89) min < med < max: 0 < 23 < 45991 IQR (CV) : 26 (9.46) | 444 distinct values |  | 3699 (25.2%) |
| CHE [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 4.79 (2.1) min < med < max: 0.98 < 4.6 < 13.89 IQR (CV) : 3.07 (0.44) | 997 distinct values |  | 2447 (16.7%) |
| AP [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 118.78 (132.87) min < med < max: 11 < 84 < 2995 IQR (CV) : 60 (1.12) | 672 distinct values |  | 1400 (9.5%) |
| LDH [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 331.15 (475.23) min < med < max: 39 < 239 < 13906 IQR (CV) : 145 (1.44) | 1137 distinct values |  | 1714 (11.7%) |
| CK [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 385.01 (2241.21) min < med < max: 8 < 80 < 98801 IQR (CV) : 142 (5.82) | 1506 distinct values |  | 2080 (14.2%) |
| GLU [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 126.41 (56.91) min < med < max: 19 < 113 < 1403 IQR (CV) : 41 (0.45) | 389 distinct values |  | 4192 (28.5%) |
| TRIG [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 141.71 (120.69) min < med < max: 14 < 115 < 5440 IQR (CV) : 82 (0.85) | 538 distinct values |  | 5061 (34.4%) |
| CHOL [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 150.8 (55.51) min < med < max: 25 < 145 < 1104 IQR (CV) : 69 (0.37) | 339 distinct values |  | 5045 (34.3%) |
| BASOR [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 0.15 (0.59) min < med < max: 0 < 0 < 23.66 IQR (CV) : 0 (4.03) | 419 distinct values |  | 732 (5.0%) |
| EOSR [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 1.3 (2.36) min < med < max: 0 < 0.59 < 73.49 IQR (CV) : 1.79 (1.82) | 927 distinct values |  | 732 (5.0%) |
| LYMR [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 14.61 (12.74) min < med < max: 0 < 11.34 < 100 IQR (CV) : 11.43 (0.87) | 3121 distinct values |  | 732 (5.0%) |
| MONOR [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 8.79 (5.81) min < med < max: 0 < 8 < 100 IQR (CV) : 5.24 (0.66) | 2334 distinct values |  | 732 (5.0%) |
| NEUR [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 75.15 (15.53) min < med < max: 0 < 78.33 < 100 IQR (CV) : 16.09 (0.21) | 3850 distinct values |  | 732 (5.0%) |
| PDW [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 12.29 (2.19) min < med < max: 6.6 < 12 < 25.3 IQR (CV) : 2.6 (0.18) | 167 distinct values |  | 1102 (7.5%) |
| RBC [numeric] | Parameter analysis value (Numeric) | Mean (sd) : 3.94 (0.77) min < med < max: 1 < 3.9 < 8.2 IQR (CV) : 1.1 (0.2) | 65 distinct values |  | 461 (3.1%) |

### 1.2.5 Suggested transformations

Next we investigate whether a pseudolog transformation of continuous variables may substantially symmetrize the univariate distributions of the continuous variables, and may hence be useful for multivariate summaries. We employ a function ida\_trans for this purpose, which optimises the parameter sigma of the pseudo-logarithm for that purpose. The optimization targets the best possible linear correlation of the transformed values with normal deviates. If no better transformation can be found, or if the improvement in correlation is less than 0.2 correlation units, no transformation is suggested.

Predictors where a transformation may symmetrize the distribution:

Number: 50

Proportion: 1

Register transformed variables in the data set:

plot side by side

### 1.2.6 Comparison of univariate distributions with and without pseudo-log transformation

The comparison is only shown for variables where a transformation is suggested.

## 1.3 Section session info

# 2. Pseudo-log transformations

## 2.1 Introduction

This supplemental section illustrates how pseudo-log transformations can be used to transform skewed distributions towards normality.

The transformation is a pseudo-logarithmic transformation mentioned by (Johnson 1949). It has the following advantages over ordinary logarithmic transformations:

* it is defined also for as
* it is a signed logarithmic transformation, and is defined also for negative values as

Of course, this comes at the cost of deviation from the logarithmic transformation in terms of interpretability.

The parameter may be used to adapt the transformation to a specific range of an empirical distribution.

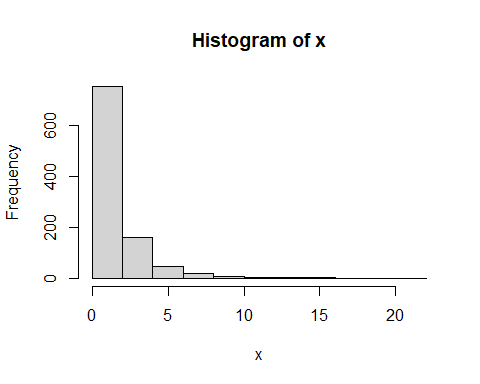
We define the pseudo-logarithmic transformation in R as:

pseudo\_log <- function(x, sigma=1, base=10) asinh(x/(2 \* sigma))/log(base)  
inv\_pseudo\_log <- function(x, sigma=1, base=10) 2 \* sigma \* sinh(x \* log(base))

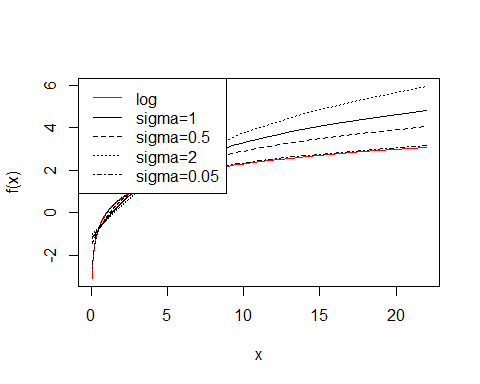
## 2.2 Impact of choice of

Next, we investigate how the parameter impacts the result of the transformation. We assume that follows a log-normal distribution, and we will show results of with different choices of . We center and scale before plotting.

p <- seq(0.001, 0.999, 0.001)  
x <- exp(qnorm(p, mean=0, sd=1))  
  
hist(x)

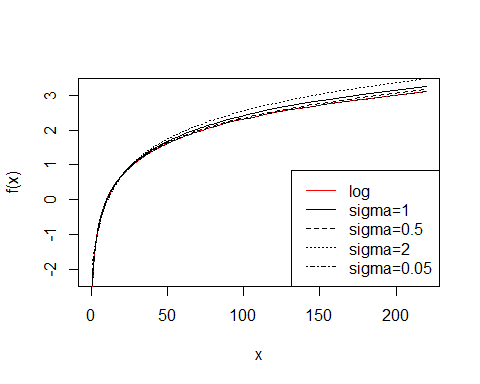


y <- cbind(log(x), scale(pseudo\_log(x, 1)), scale(pseudo\_log(x,0.5)), scale(pseudo\_log(x, 2)), scale(pseudo\_log(x, 0.05)))  
  
  
plot(x, y[,2], type="l", ylab="f(x)", ylim=range(y))  
lines(x, y[,1], col="red")  
lines(x, y[,3], type="l", lty=2)  
lines(x, y[,4], type="l", lty=3)  
lines(x, y[,5], type="l", lty=4)  
legend("topleft", lty=c(1,1,2,3,4), col=c("red","black","black","black","black"),   
 legend=c("log", "sigma=1", "sigma=0.5", "sigma=2", "sigma=0.05"))



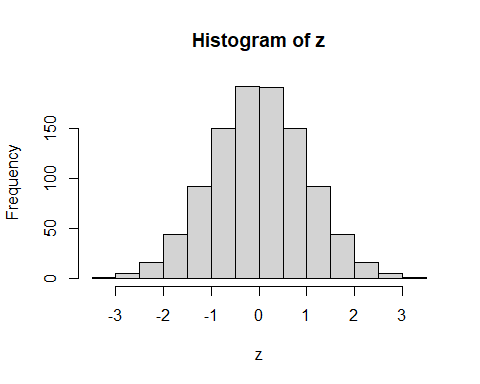
In the next code chunk, we multiply by 10 and repeat the exercise. We learn that the transformations become more similar and the choice of less relevant.

x <- x\*10  
  
y <- cbind(scale(log(x)), scale(pseudo\_log(x, 1)), scale(pseudo\_log(x,0.5)), scale(pseudo\_log(x, 2)), scale(pseudo\_log(x, 0.05)))  
  
  
plot(x, y[,2], type="l", ylab="f(x)")  
lines(x, y[,1], col="red")  
lines(x, y[,3], type="l", lty=2)  
lines(x, y[,4], type="l", lty=3)  
lines(x, y[,5], type="l", lty=4)  
legend("bottomright", lty=c(1,1,2,3,4), col=c("red","black","black","black","black"),   
 legend=c("log", "sigma=1", "sigma=0.5", "sigma=2", "sigma=0.05"))

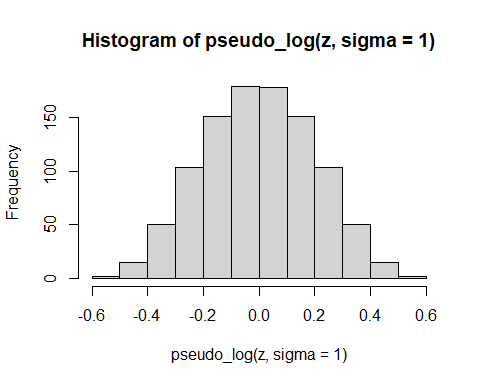


Finally, we apply the pseudo-logarithmic transformation to the original normal deviates. We learn that a higher value for the parameter makes the distribution ‘slimmer’ while a lower value makes it ‘fatter’, and it is even possible to induce bimodality with low values of sigma:

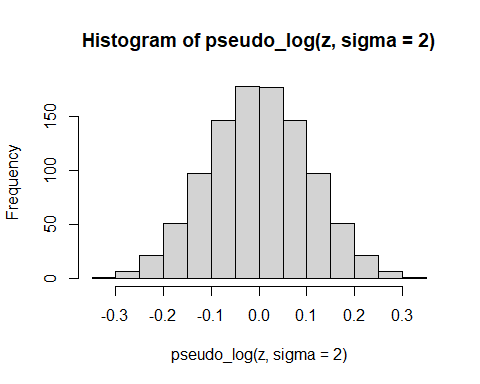
z <- qnorm(p, mean=0, sd=1)  
hist(z)



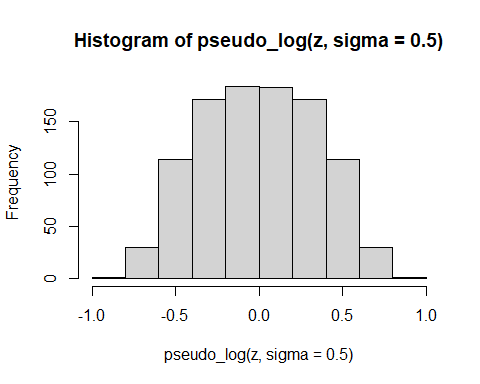
hist(pseudo\_log(z, sigma=1))



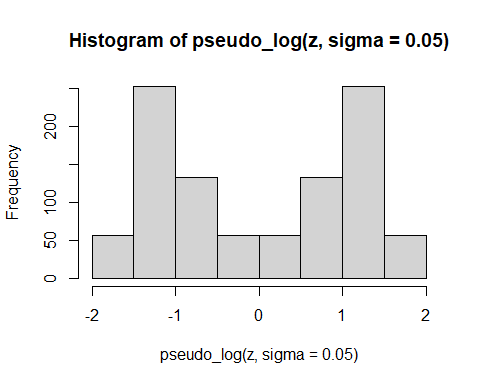
hist(pseudo\_log(z, sigma=2))



hist(pseudo\_log(z, sigma=0.5))



hist(pseudo\_log(z, sigma=0.05))



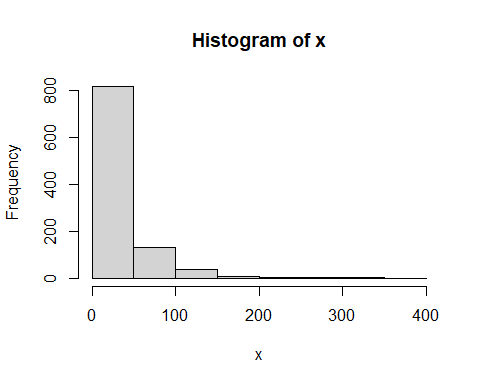
## 2.3 Finding a parameter that best achieves normality

Any test statistic for testing normality could be chosen to find a suitable value of that induces normality into the transformed values. Here we use the (Pearson) correlation coefficient to compare the empirical distribution with normal deviates.

### 2.3.1 Deviates from log normal distribution

We simulate from a shifted log normal distribution and evaluate the value of that optimizes agreement with a normal distribution:

x<-sort(exp(rnorm(1000)+3))  
  
hist(x)



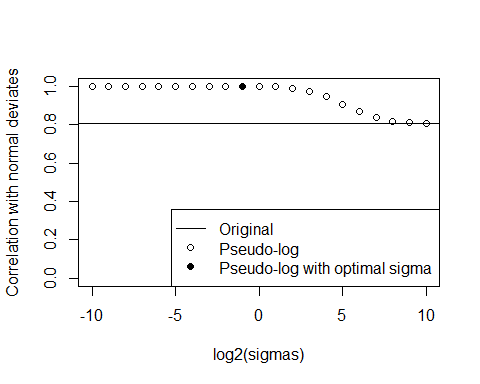
sigmas <- 2\*\*seq(-10, 10, 1)  
origcor <- cor(qnorm((1:length(x)-0.5)/length(x)), x)  
  
ncorx <- sapply(sigmas, function(X) cor(qnorm((1:length(x)-0.5)/length(x)), pseudo\_log(x,X)))  
  
cat("Optimal sigma: ")

Optimal sigma:

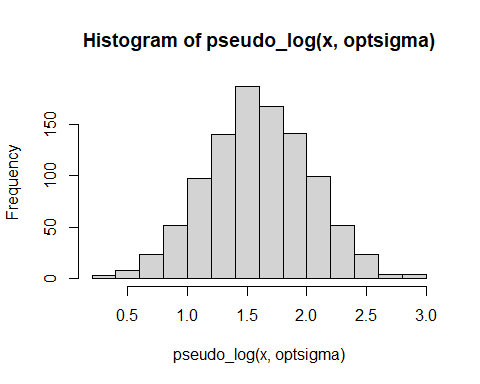
(optsigma<-sigmas[ncorx==max(ncorx)])

[1] 0.5

plot(log2(sigmas), ncorx, ylab="Correlation with normal deviates", ylim=c(0,1))  
points(log2(sigmas)[ncorx==max(ncorx)], max(ncorx), pch=19)  
abline(h=origcor)  
legend("bottomright", lty=c(1, NA,NA), pch=c(NA,1,19), legend=c("Original", "Pseudo-log", "Pseudo-log with optimal sigma"))  
box()



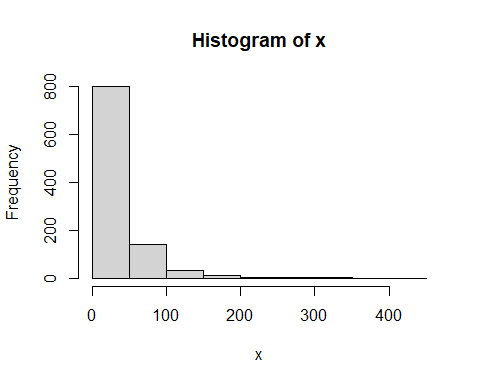
hist(pseudo\_log(x, optsigma))



### 2.3.2 Deviates from exponential distribution

Also with an exponential distribution, the pseudo-logarithm may achieve a better agreement with a normal:

x<-sort(exp(rnorm(1000)+3))  
  
hist(x)



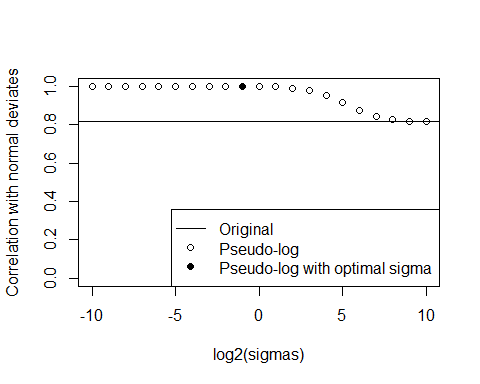
sigmas <- 2\*\*seq(-10, 10, 1)  
origcor <- cor(qnorm((1:length(x)-0.5)/length(x)), x)  
  
ncorx <- sapply(sigmas, function(X) cor(qnorm((1:length(x)-0.5)/length(x)), pseudo\_log(x,X)))  
  
cat("Optimal sigma: ")

Optimal sigma:

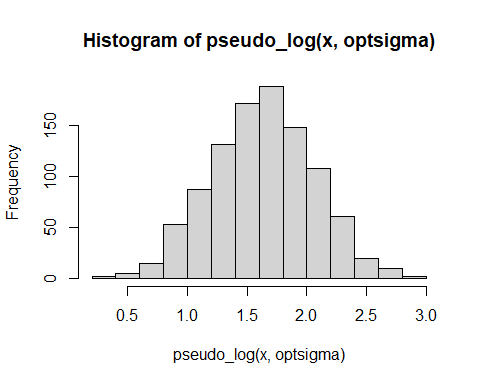
(optsigma<-sigmas[ncorx==max(ncorx)])

[1] 0.5

plot(log2(sigmas), ncorx, ylab="Correlation with normal deviates", ylim=c(0,1))  
points(log2(sigmas)[ncorx==max(ncorx)], max(ncorx), pch=19)  
abline(h=origcor)  
legend("bottomright", lty=c(1, NA,NA), pch=c(NA,1,19), legend=c("Original", "Pseudo-log", "Pseudo-log with optimal sigma"))  
box()



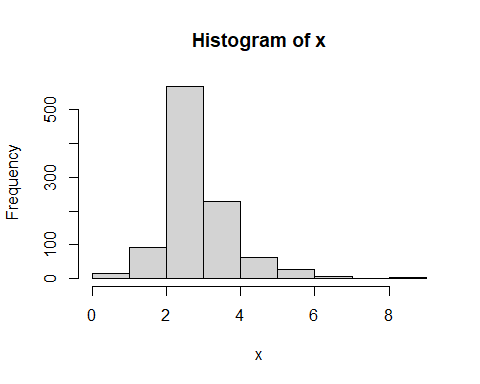
hist(pseudo\_log(x, optsigma))



### 2.3.3 Deviates from mixture distribution

Now we mix a normal, lognormal and exponential distribution:

x1 <- scale(rnorm(1000))  
x2 <- scale(rexp(1000))  
x3 <- scale(exp(rnorm(1000)))  
  
p1 <- rbinom(1000, 1, 0.33)  
p2 <- rbinom(1000, 1, 0.5)  
  
x<-p1\*x1 + (1-p1)\*(p2\*x2+(1-p2)\*x3)  
x <- x-min(x)  
x<-sort(x)  
  
hist(x)



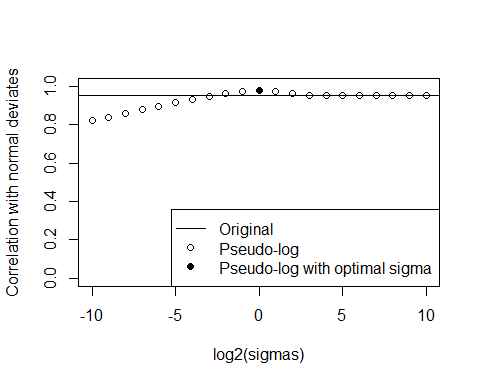
sigmas <- 2\*\*seq(-10, 10, 1)  
origcor <- cor(qnorm((1:length(x)-0.5)/length(x)), x)  
  
ncorx <- sapply(sigmas, function(X) cor(qnorm((1:length(x)-0.5)/length(x)), pseudo\_log(x,X)))  
  
cat("Optimal sigma: ")

Optimal sigma:

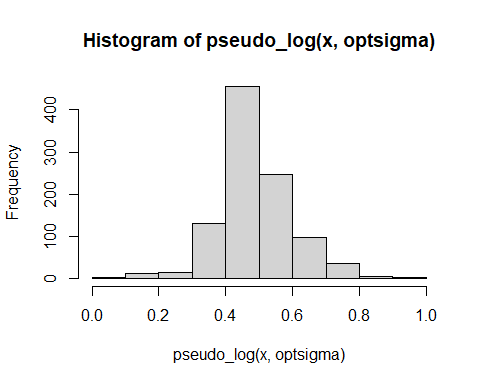
(optsigma<-sigmas[ncorx==max(ncorx)])

[1] 1

plot(log2(sigmas), ncorx, ylab="Correlation with normal deviates", ylim=c(0,1))  
points(log2(sigmas)[ncorx==max(ncorx)], max(ncorx), pch=19)  
abline(h=origcor)  
legend("bottomright", lty=c(1, NA,NA), pch=c(NA,1,19), legend=c("Original", "Pseudo-log", "Pseudo-log with optimal sigma"))  
box()



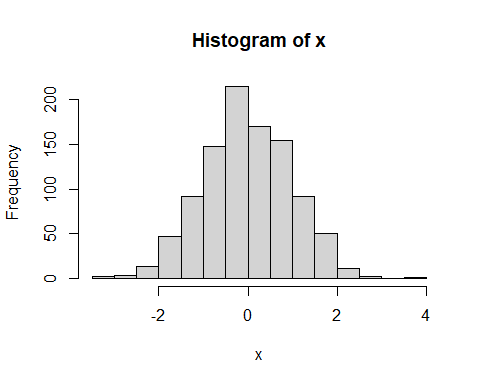
hist(pseudo\_log(x, optsigma))



### 2.3.4 Normal deviates

With simulated normal deviates, the pseudo-logarithm cannot improve the already perfect normality.

x<-sort(rnorm(1000))  
  
hist(x)



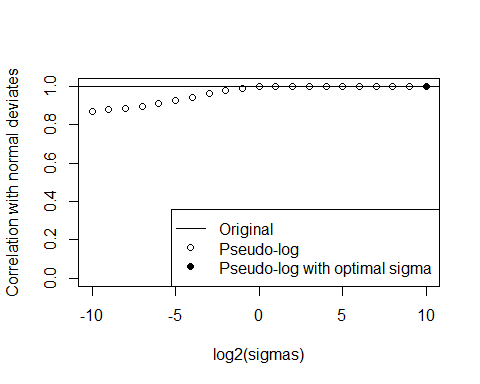
sigmas <- 2\*\*seq(-10, 10, 1)  
origcor <- cor(qnorm((1:length(x)-0.5)/length(x)), x)  
  
ncorx <- sapply(sigmas, function(X) cor(qnorm((1:length(x)-0.5)/length(x)), pseudo\_log(x,X)))  
  
cat("Optimal sigma: ")

Optimal sigma:

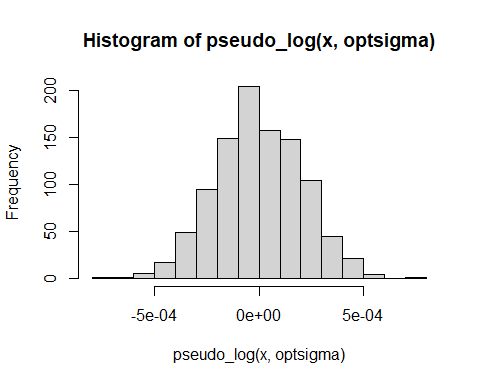
(optsigma<-sigmas[ncorx==max(ncorx)])

[1] 1024

plot(log2(sigmas), ncorx, ylab="Correlation with normal deviates", ylim=c(0,1))  
points(log2(sigmas)[ncorx==max(ncorx)], max(ncorx), pch=19)  
legend("bottomright", lty=c(1, NA,NA), pch=c(NA,1,19), legend=c("Original", "Pseudo-log", "Pseudo-log with optimal sigma"))  
abline(h=origcor)  
box()



hist(pseudo\_log(x, optsigma))



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>.