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. Supplementary Example

## 1.1 Overview

Define key predictors without pseudo-log transform (‘orig’) and transformed (‘trans’).

In the following examples we use the Bacteremia data with complete observations regarding the key predictors PLT, BUN, NEU, AGE, CREA\_T, WBC\_noNEU\_T, which represent 93.9% of the dataset **w.r.t key predictors**. We will fit a global logistic regression model with the outcome **BACTEREMIAN** (i.e. presence of bacteremia) and the key predictors as covariates. We will use pseudo-log transformations as suggested in the IDA. Within the model, all key predictors will be transformed by fractional polynomials of order 1 (df = 2).

The aim of the examples is to showcase how decisions derived from IDA influence the results of the fitted model.

## 1.2 Global Model

The global model will be fitted by the *mfp* function. If not indicated otherwise, we will use the fp-transformations of the key predictors determined in global model in all consecutive models. For all models we report McFaddens’s R² and the AUC, i.e. the area under the ROC curve, and boxplots comparing “**BACTEREMIAN**” predictions with outcomes.

### 1.2.1 Model Summary

The global model is specified as print(global\_formula, quote = TRUE). The model is fit to the complete cases data on the **transformed** scale stored in the data set data\_model\_trans.

Report the global model fit.

Table 1: global model - including transformed variables

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | -2.70 | 0.59 | -4.59 | 4.44 × 10^-6 |
| I((WBC\_noNEU\_T + 0.1)^0.5) | -5.23 | 0.25 | -20.51 | 1.90 × 10^-93 |
| I(((NEU + 0.1)/10)^0.5) | 1.52 | 0.12 | 13.08 | 4.05 × 10^-39 |
| I((AGE/100)^1) | 1.69 | 0.20 | 8.48 | 2.18 × 10^-17 |
| I(CREA\_T^1) | 0.71 | 0.20 | 3.59 | 3.25 × 10^-4 |
| I(((PLT + 1)/100)^1) | -0.08 | 0.03 | -2.57 | 1.03 × 10^-2 |
| I((BUN/10)^1) | 0.01 | 0.02 | 0.43 | 6.64 × 10^-1 |

Plot predictions by outcome and report the model AUC.

|  |
| --- |
| Global model - including covariates on the transformed psuedo log scale |

### 1.2.2 Functional forms of global model

We now take a look at the functional forms of the covariates in the global model, which are determined by the fp algorithm. Besides scaling factors, only for WBC\_noNEU\_T the fp algorithm chose a non-linear transformation (note the ‘^0.5’ in the term column). This means all other covariates enter the model in a linear fashion. In the following effect plots, each variable is adjusted to the median of the other variables in the model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  | | --- | | (a) | | |  | | --- | | (b) | | |  | | --- | | (c) | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| |  | | --- | | (d) | | |  | | --- | | (e) | | |  | | --- | | (f) | |

Figure 1.1: Functional forms of global model - repeats over plots

## 1.3 Example 1: to transform or not to transform

Only for one out of the six key predictors did the fp algorithm chose a non-linear transformation. But out of those six variables, four were pseudo-log transformed before entering the model. In the first example we want to compare the global model to a model using the key predictors on their original scale.

Table 1: Global model without pseudo-log tranformations

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | -3.18 | 0.28 | -11.17 | 5.48 × 10^-29 |
| log((WBC\_noNEU + 0.2)) | -1.25 | 0.06 | -20.22 | 5.92 × 10^-91 |
| I(((NEU + 0.1)/10)^0.5) | 1.52 | 0.12 | 12.99 | 1.34 × 10^-38 |
| I((AGE/100)^1) | 1.64 | 0.20 | 8.26 | 1.51 × 10^-16 |
| I(((PLT + 1)/100)^1) | -0.08 | 0.03 | -2.53 | 1.13 × 10^-2 |
| I((BUN/10)^1) | 0.02 | 0.02 | 0.68 | 5.00 × 10^-1 |
| I(CREA^-0.5) | -0.77 | 0.21 | -3.69 | 2.20 × 10^-4 |

Note the different fp-transformation arising when the key predictors are not pseudo-log transformed. On the original scale, three covariates instead of one now enter the model via a non-linear fp-transformation. This suggests that a transformation prior to the regression model ‘outsources’ the need for transformations within the model. Now let us compare the model performances.

|  |
| --- |
| Global model(s) with covariates on original and transformed scale. |

With regards to McFadden’s R² and the AUC, the differences between the two approaches is marginal.

Next we will compare the differences of the functional forms in the two models for those covariates where a pseudo-log transformation was suggested in IDA. We will look at the log odds for bacteremia by each covariate on the original and the transformed scale, and compare the global model using the original and the pseudo-log transformed covariates. Each variable is adjusted for the median of all other variables used.

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

|  |
| --- |
| Comparing functional form of both global models. |

## 1.4 Example 2: the support of a model determines what it can explain

Next we compare the global model to a model were for an important variable, in our case we chose age, the variable support is reduced to the central 50% of the data (i.e. data within the 25% and 75% quantiles). Again, in the reduced models we use the same fp-transformations as in the global model.

| data | AUC | scaled Brier score |
| --- | --- | --- |
| complete | 0.731 | 0.06485 |
| central 50% | 0.722 | 0.05288 |
| 50% sample | 0.717 | 0.05625 |

|  |
| --- |
| Model predictions based on different missing data scenarios. |

## 1.5 Example 3: the limits of mulitiple imputation

To show the effect of multiple imputation if the number of missing values is high, we construct a dataset with 50% artificially generated missing values in one variable. First, recall the output of the complete model, relying on the Bacteremia data with complete cases regarding the key predictors.

Table 1: global model

| term | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- |
| (Intercept) | -2.70 | 0.59 | -4.59 | 4.44 × 10^-6 |
| I((WBC\_noNEU\_T + 0.1)^0.5) | -5.23 | 0.25 | -20.51 | 1.90 × 10^-93 |
| I(((NEU + 0.1)/10)^0.5) | 1.52 | 0.12 | 13.08 | 4.05 × 10^-39 |
| I((AGE/100)^1) | 1.69 | 0.20 | 8.48 | 2.18 × 10^-17 |
| I(CREA\_T^1) | 0.71 | 0.20 | 3.59 | 3.25 × 10^-4 |
| I(((PLT + 1)/100)^1) | -0.08 | 0.03 | -2.57 | 1.03 × 10^-2 |
| I((BUN/10)^1) | 0.01 | 0.02 | 0.43 | 6.64 × 10^-1 |

Creatinine (‘CREA’) is significant at a level that might not survive substantial missingness. We thus create a dataset were we artificially introduce 50% missing creatinine values, missing completely at random.

Next we fit a ‘complete case’ model in the case of missing creatinine data, using the fp-transformations from the global model.

Now we impute the missing creatinine data using MICE with 50 imputations, fit the model using the fp-transformations from the global model and pool the results.

We now can compare the outputs of the complete model, the complete model with missing data (i.e. only half of the original complete data is used), and the imputed model.

Table 1: Comparison of the complete model, the complete model with missing data

| term | model | estimate | std.error | statistic | p.value |
| --- | --- | --- | --- | --- | --- |
| (Intercept) | global model | -2.698 | 0.588 | -4.590 | 4.44 × 10^-6 |
|  | missing, complete cases | -2.019 | 0.841 | -2.401 | 1.64 × 10^-2 |
|  | missing, imputed | -1.765 | 0.787 | -2.243 | 2.59 × 10^-2 |
| I(((NEU + 0.1)/10)^0.5) | global model | 1.525 | 0.117 | 13.084 | 4.05 × 10^-39 |
|  | missing, complete cases | 1.579 | 0.168 | 9.387 | 6.20 × 10^-21 |
|  | missing, imputed | 1.523 | 0.117 | 13.073 | 0.00 |
| I(((PLT + 1)/100)^1) | global model | -0.082 | 0.032 | -2.565 | 1.03 × 10^-2 |
|  | missing, complete cases | -0.042 | 0.045 | -0.946 | 3.44 × 10^-1 |
|  | missing, imputed | -0.081 | 0.032 | -2.549 | 1.08 × 10^-2 |
| I((AGE/100)^1) | global model | 1.688 | 0.199 | 8.484 | 2.18 × 10^-17 |
|  | missing, complete cases | 2.160 | 0.284 | 7.599 | 2.99 × 10^-14 |
|  | missing, imputed | 1.674 | 0.199 | 8.409 | 0.00 |
| I((BUN/10)^1) | global model | 0.010 | 0.024 | 0.434 | 6.64 × 10^-1 |
|  | missing, complete cases | 0.033 | 0.034 | 0.987 | 3.24 × 10^-1 |
|  | missing, imputed | 0.039 | 0.029 | 1.371 | 1.71 × 10^-1 |
| I((WBC\_noNEU\_T + 0.1)^0.5) | global model | -5.225 | 0.255 | -20.506 | 1.90 × 10^-93 |
|  | missing, complete cases | -5.354 | 0.366 | -14.648 | 1.39 × 10^-48 |
|  | missing, imputed | -5.208 | 0.255 | -20.452 | 0.00 |
| I(CREA\_T^1) | global model | 0.712 | 0.198 | 3.595 | 3.25 × 10^-4 |
|  | missing, complete cases | 0.361 | 0.280 | 1.289 | 1.97 × 10^-1 |
|  | missing, imputed | 0.374 | 0.272 | 1.376 | 1.70 × 10^-1 |

The z-statistic for creatinine drops from 2.98 to 1.49 when half the data is missing. Also in other variables the z-statistic is less extreme in the ‘missing, complete case analysis’ compared to the global model. The interesting observations is that MI recreates estimates and standard errors very close to the global model in most variables, but not in the one that was being imputed, namely creatinine. In variable selection, creatinine, which is highly significant in the ‘true’ model, is likely to be dropped, based on the imputed data.

## 1.6 Example 4: Interpretation of regression coefficient ‘size’

The variables WBC\_noNEU and WBC\_noNEU\_T are on two very different scales:

Plot

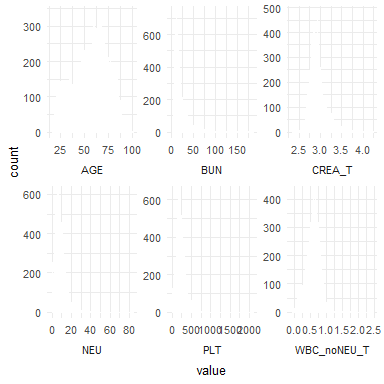


Table the standardized regression coefficients.

Table 1: Global model - standardized regression estimates

| term | estimate |
| --- | --- |
| I((WBC\_noNEU\_T + 0.1)^0.5) | -5.225 |
| I(((NEU + 0.1)/10)^0.5) | 1.525 |
| I((AGE/100)^1) | 1.688 |
| I(CREA\_T^1) | 0.712 |
| I(((PLT + 1)/100)^1) | -0.082 |
| I((BUN/10)^1) | 0.010 |

Table standardized coefficients.

Table 1: Global model - standardized regression coefficients

| variable | standardized beta |
| --- | --- |
| WBC\_noNEU | 0.7029 |
| NEU | 8.5596 |
| AGE | 0.3071 |
| CREA | 0.1561 |
| PLT | 0.0993 |
| BUN | 0.1807 |

Let us recall the two estimates to the covariates WBC\_noNEU and WBC\_noNEU\_T.

Table 1: Comparing original scale and transformed

| term | estimate |
| --- | --- |
| I((WBC\_noNEU\_T + 0.1)^0.5) | -5.23 |
| log((WBC\_noNEU + 0.2)) | -1.25 |

(Suggestion: show this with models without fp transform?)

Because the fp-transformations further complicate the interpretation of the regression coefficients, let us consider two logisitc regression models with WBC\_noNEU and t\_WBC\_noNEU as single covariate, respectively.

Table 1: Comparing original scale and transformed

| term | estimate |
| --- | --- |
| WBC\_noNEU | -0.56 |
| WBC\_noNEU\_T | -2.89 |

The estimates -0.56 and -2.89 the change in log odds for the outcome when the ‘term’ variable changes by 1 unit, but cannot be compared directly. A unit change is only a small step on the original scale, where WBC\_noNEU covers values from -0.15 up to 152.74. In comparison, WBC\_noNEU\_T lies between up to -, so change of unit cover almost half the range of the variable.

## 1.7 Example 5: Plot of functional form should be resticted to areas with high density

The functional forms have wide confidence intervals when the data is sparse. In presentations of the effects, plots of the functional forms can be limited to areas with high density. In this analysis, PLT was very sparse above ~800 [UNITS], which is reflected in a large confidence interval for high PLT values. In the effect plot PLT values could be limited to values <800 [UNITS].

|  |
| --- |
| Model predictions |

|  |
| --- |
| TODO: Check this plots |

|  |
| --- |
| PLT comparison |