Weight of Evidence Regression for Probability of Default Modeling

Intercept Estimation Uncertainty

Andrija Djurovic

www.linkedin.com/in/andrija-djurovic

Probability of Default Modeling in Practice

- For the development of Probability of Default (PD) models, logistic regression remains the most commonly used statistical method.
- Practitioners often choose to discretize all numeric risk factors used in logistic regression.
- Discretization and the use of categorical risk factors require a specific encoding method.
- The most common encoding method in PD modeling with logistic regression is Weight of Evidence (WoE).
- When WoE is applied in logistic regression, practitioners usually refer to the model as WoE regression.
- The main advantages of WoE encoding in logistic regression are its natural fit, since it is expressed in terms of log-odds, and its interpretability, as multiplying it by the regression coefficient shows the additional contribution of a specific risk factor modality to the target level.
- The additional contribution of a risk factor modality in WoE regression is tied to the role of the intercept. The intercept in WoE regression has a specific interpretation. In the context of PD modeling, it represents the overall default rate.
- Given this interpretation of the intercept in WoE regression, this presentation investigates the uncertainty of intercept estimation under the Iteratively Reweighted Least Squares (IWLS) method for coefficient estimation in logistic regression. The following slide presents the main concept and WoE calculation, while the subsequent slides present a simulation study of intercept estimation uncertainty in WoE regression.

Weights of Evidence Encoding

The Weight of Evidence (WoE) for modality i of a risk factor is defined in terms of the distributions of goods (non-defaulters) and bads (defaulters) as:

$$\text{WoE}_i = \ln\left(\frac{\text{Dist}_i^{\text{Good}}}{\text{Dist}_i^{\text{Bad}}}\right)$$

where:

- $\text{Dist}_i^{\text{Good}} = \dfrac{N_i^{\text{Good}}}{N^{\text{Good}}}$ is the of all goods (non-defaulters) that fall into category i;
- $\text{Dist}_i^{\text{Bad}} = \dfrac{N_i^{\text{Bad}}}{N^{\text{Bad}}}$ is the proportion of all bads (defaulters) that fall into category i;
- N_i^{Good} is the number of goods (non-defaulters) in category i;
- N_i^{Bad} is the number of bads (defaulters) in category i;
- N^{Good} is the total number of goods (non-defaulters) in the sample;
- N^{Bad} is the total number of bads (defaulters) in the sample.

Simulation Study

Dataset

The modeling dataset used for this simulation consists of the target variable and ten risk factors and is available at the following link. The column `Creditability` contains the target variable, while the remaining columns represent the risk factors used in the simulation.

Simulation Design

The simulation begins by generating all possible combinations of five to ten risk factors, presented in the order specified in the dataset used for the simulation. For each combination, a logistic regression with WoE-encoded risk factors is estimated, and the intercept value is collected. The intercept log-odds are then transformed into probabilities. Finally, the basic descriptive statistics and a visualization of the simulated intercept values are presented.

The following slides present the simulation results.

Simulation Results

The following table summarizes the simulation results, showing the basic descriptive statistics of the simulated intercept, transformed into probabilities, for different numbers of risk factors used in the WoE regression.

Number of Risk Factors	# of Models	Minimum	Average	Median	Maximum	Standard Deviation
5	1287	0.2972	0.2999	0.2998	0.3038	0.0009
6	1716	0.2967	0.2998	0.2998	0.3038	0.0010
7	1716	0.2964	0.2997	0.2997	0.3037	0.0011
8	1287	0.2964	0.2996	0.2996	0.3039	0.0012
9	715	0.2968	0.2994	0.2994	0.3031	0.0012
10	286	0.2967	0.2992	0.2992	0.3023	0.0011

Simulation Results cont.