Supplementary Material to the paper 'Network Model with Application to Allergy Diseases'

Konrad Furmańczyk[1,5][0000−0002−7683−4787], Wojciech Niemiro[2,3][0000−0002−7076−8838], Mariola Chrzanowska[4,5][0000−0002−8743−7437], and Marta Zalewska[5][0000−0002−8163−961X]

[1] Institute of Information Technology, Warsaw University of Life Sciences, Warsaw, Poland konrad_furmanczyk@sggw.edu.pl
[2] Faculty of Mathematics, Informatics and Mechanics University of Warsaw, Poland
[3] Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Poland wniemiro@gmail.com
[4] Institute of Economics and Finance, Warsaw University of Life Sciences, Poland mariola_chrzanowska@sggw.edu.pl
[5] Department of Prevention of Environmental Hazards, Allergology and Immunology, Medical University of Warsaw, Poland marta.zalewska@wum.edu.pl

A Estimation for the misspecified model

A.1 The Weighted Logistic Regression

We do not always have appropriate sample sizes for rare diseases and work with imbalanced datasets. In such cases, we may improve prediction accuracy for logistic regression using weighted logistic regression ([3], [5]) or apply a machine learning algorithm such as use SMOTE Simple Genetic Algorithm ([2]) to determine the sampling rate of each example in order to get unequal synthetic samples or using undersampling or oversampling ([4]). However, resampling techniques do not easily transfer to dependent logistic regression equations. For this reason, in the paper, we use weighted regression as in [3]. Following the approach of [3] we penalized misclassification costs of events and non-events differently by penalty weights w_1 and w_0 in the log-likelihood function for each i equation

$$min_{\theta_i} \left\{ -w_1 \sum_{j=1}^{n} y_{ij} log(\sigma(\mathbf{z}_j^T \theta_i)) - w_0 \sum_{j=1}^{n} (1 - y_{ij}) log(1 - \sigma(\mathbf{z}_j^T \theta_i)) \right\},$$

where n is a sample size, $w_1 = \frac{\tau_i}{\bar{y_i}}$ and $w_0 = \frac{1-\tau_i}{1-\bar{y_i}}$, and τ_i denoting the population fraction of events induced by choice-based sampling and $\bar{y_i}$ denoting the sample proportion of events, θ_i is a vector of all parameters, $\mathbf{z}_j$ is a vector of all predictors, and $\sigma(x) = \frac{exp(x)}{1+exp(x)}$.

Model estimation is performed separately for each equation (see formula (4)-(5) in the main paper) using the standard GLM procedure for logistic regression in the first scenario, and in the second scenario, we use weighted logistic regression.

According to work of [1] we assume that the population fraction in Poland for considered allergy diseases are as follows $\tau_1 = 11\%, \tau_2 = 20\%, \tau_3 = 4\%, \tau_4 = 7\%, \tau_5 = 10\%$.

We applied weigthed logistic regression to the EACP data in parallel with the standard unweighted GLM. It turned out that both methods gave very similar results (see A.2-A.3).

A.2 Results for the standard logistic regression

The standard errors for standard logistic regression coefficients estimation are given in Tables 1-2. Next, we present the odds ratio with the asymptotic 0.95 confidence interval (CI) for the standard logistic regression (see Tables 3-6).

Table 1. The standard errors of estimation for standard logistic regression - Part 1

$logit_i$	ω_{0i}	α_{1i}	α_{2i}	α_{3i}	α_{4i}	β_{1i}	β_{2i}	β_{3i}	β_{4i}	β_{5i}
i=1	0.350	0.197	0.190	0.219	0.149	0.206	0.228	0.180	0.304	0.304
i=2	0.212	0.121	0.113	0.133	0.088	0.123	0.134	0.107	0.184	0.239
i=3	0.223	0.116	0.110	0.142	0.086	0.117	0.130	10.110	0.185	0.224
i=4	0.393	0.145	0.163	0.244	0.127	0.150	0.169	0.151	0.243	0.260

Table 2. The standard errors of estimation for standard logistic regression - Part 2

$logit_i$	γ_{i1}	γ_{i2}	γ_{i3}	ω_{2i}	ω_{3i}	ω_{4i}	ω_{5i}
i=1	0.151	-	-	0.184	0.158	0.222	-
i=2	-	0.096	-	-	-	0.164	-
i=3	-	0.096	-	-	-	0.148	-
i=4	-	-	0.167	-	-	-	0.147

Table 3. The OR with 0.95 CI for estimation results for standard logistic regression - Part 1

$logit_i$	$\exp(\alpha_{1i})$	$\exp(\alpha_{2i})$	$\exp(\alpha_{3i})$	$\exp(\alpha_{4i})$
i=1	1.426(0.969;2.098)	1.121(0.855;1.801)	0.715(0.465;1.098)	1.510(1.127;2.022)
i=2	1.332(1.051;1.689)	1.439(1.153;1.796)	0.681(0.525;0.884)	0.963(0.810;1.144)
i=3	1.398(1.114;1.755)	1.311(1.057;1.627)	0.996(0.754;1.316)	1.397(1.180;1.653)
i=4	1.215(0.915;1.615)	0.536(0.390;0.738)	1.484(0.920;2.395)	0.902(0.703;1.157)

Table 4. The OR with 0.95 CI for estimation results for standard logistic regression - Part 2

$logit_i$	$\exp(\beta_{1i})$	$\exp(\beta_{2i})$	$\exp(\beta_{3i})$	$\exp(\beta_{i4})$	$\exp(\beta_{i5})$
i=1	0.928(0.620;1.389)	0.898(0.574;1.403)	1.214(0.853;1.728)	1.029(0.567;1.868)	2.088(1.150;3.788)
i=2	0.956(0.751;1.217)	1.328(1.022;1.727)	1.368(1.109;1.687)	1.094(0.763;1.569)	0.874(0.547;1.396)
i=3	1.057(0.840;1.329)	1.204(0.934;1.554)	0.962(0.775;1.193)	0.866(0.603;1.244)	0.978(0.631;1.517)
i=4	1.392(1.038;1.868)	1.246(0.895;1.735)	1.063(0.791;1.429)	0.684(0.425;1.101)	1.600(0.961;2.663)

Table 5. The OR with 0.95 CI for estimation results for standard logistic regression - Part 3

$logit_i$	$\exp(\gamma_{i1})$	$\exp(\gamma_{i2})$	$\exp(\gamma_{i3})$
i=1	4.104(3.053;5.518)	-	-
i=2	-	4.015(3.326;4.846)	-
i=3	-	5.094(4.220;6.148)	-
i=4	-	-	5.930(4.275;8.226)

Table 6. The OR with 0.95 CI for estimation results for standard logistic regression - Part 4

$logit_i$	$\exp(\omega_{2i})$	$\exp(\omega_{3i})$	$\exp(\omega_{4i})$	$\exp(\omega_{5i})$
i=1	3.543(2.470;5.082)	7.691(5.642;10.482)	2.040(1.320;3.152)	-
i=2	-	-	1.154(0.837;1.591)	-
i=3	-	-	1.642(1.229;2.195)	-
i=4	-	-	-	3.102(2.325;4.138)

A.3 Results for the weighted logistic regression

The results for estimation are given in Tables 7-8. Next, we present the odds ratio with the asymptotic 0.95 confidence interval (CI) for the weighted logistic regression (see Tables 9-10).

Table 7. Estimation results for weighted logistic regression - Part 1

$logit_i$	ω_{0i}	α_{1i}	α_{2i}	α_{3i}	α_{4i}	β_{1i}	β_{2i}	β_{3i}	β_{4i}	β_{5i}
i=1	-5.053	0.356	0.206	-0.349	0.447	-0.039	-0.092	0.181	0.024	0.737
i=2	-3.544	0.289	0.355	-0.377	-0.042	-0.041	0.285	0.310	0.088	-0.154
i=3	-6.212	0.319	0.263	0.024	0.341	0.061	0.186	-0.042	-0.163	0.001
i=4	-3.301	0.139	-0.646	0.278	-0.104	0.414	0.322	-0.009	-0.480	0.409

Table 8. Estimation results for weighted logistic regression - Part 2

$logit_i$	γ_{i1}	γ_{i2}	γ_{i3}	ω_{2i}	ω_{3i}	ω_{4i}	ω_{5i}
i=1	1.392	-	-	1.392	2.085	0.753	-
i=2	-	1.377	-	-	-	0.140	-
i=3	-	1.628	-	-	-	0.450	-
i=4	-	-	1.800	-	-	-	1.302

Table 9. The OR for estimation results for weighted logistic regression - Part 1

$logit_i$	$\exp(\alpha_{1i})$	$\exp(\alpha_{2i})$	$\exp(\alpha_{3i})$	$\exp(\alpha_{4i})$	$\exp(\beta_{1i})$	$\exp(\beta_{2i})$	$\exp(\beta_{3i})$	$\exp(\beta_{4i})$	$\exp(\beta_{5i})$
i=1	1.428	1.229	0.705	1.564	0.962	0.912	1.198	1.024	2.090
i=2	1.335	1.426	0.686	0.959	0.960	1.330	1.363	1.092	0.857
i=3	1.376	1.301	1.024	1.406	1.063	1.204	0.959	0.850	1.001
i=4	1.149	0.524	1.320	0.901	1.513	1.380	0.991	0.619	1.505

Table 10. The OR for estimation results for weighted logistic regression - Part 2

$logit_i$	$\exp(\gamma_{i1})$	$\exp(\gamma_{i2})$	$\exp(\gamma_{i3})$	$\exp(\omega_{2i})$	$\exp(\omega_{3i})$	$\exp(\omega_{4i})$	$\exp(\omega_{5i})$
i=1	4.023	-	-	4.023	8.045	2.123	-
i=2	-	3.963	-	-	-	1.150	-
i=3	-	5.094	-	-	-	1.568	-
i=4	-	-	6.050	-	-	-	3.677

B Evaluation of the misspecified model

The ROC curve for $logit_2 - logit_4$ for bootstrap, jackknofe for weighted and unweighted estimation we present in Figures 1-15. In general, the method without weights gave better AUC results except in the case of $logit_4$. However, the differences were quite negligible. When the jackknife method was used, also clear difference was not observed. In general, weighting did not really help in estimating model parameters, except for the equation for $logit_4$.

Table 11. AUC for each logit

$logit_i$	*weight*	*boot+weight*	*boot*	*jackkn+weight*	*jackkn*
i=1	0.8406	0.8258	0.8470	0.8231	0.8165
i=2	0.6704	0.6907	0.6986	0.6700	0.6857
i=3	0.7234	0.7196	0.7201	0.7259	0.7215
i=4	0.7971	0.7936	0.7931	0.7861	0.7921

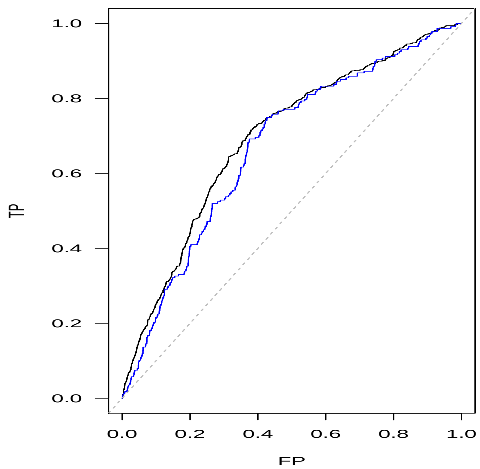

Fig. 1. ROC for $logit_2$ for unweighted (black curve) and weigthed (blue curve) estimation.

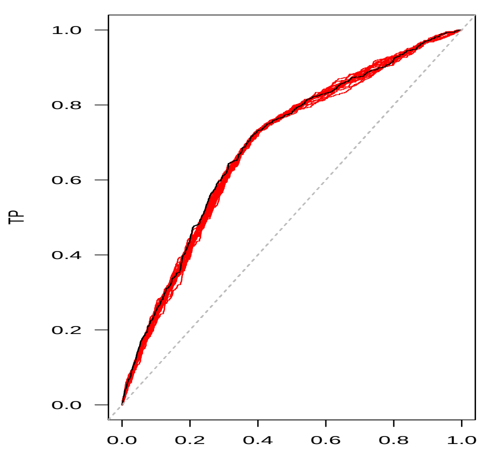

Fig. 2. ROC for $logit_2$ for unweighted estimation-bootstrap.

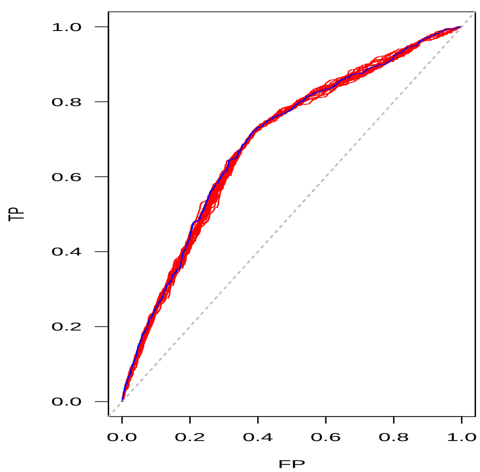

Fig. 3. ROC for $logit_2$ for weighted estimation-bootstrap.

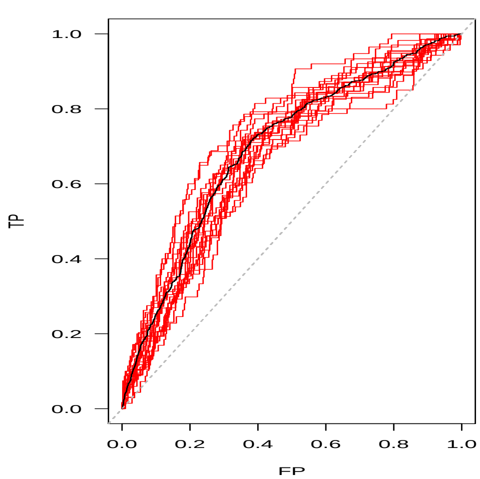

Fig. 4. ROC for $logit_2$ for unweighted estimation-jackknife.

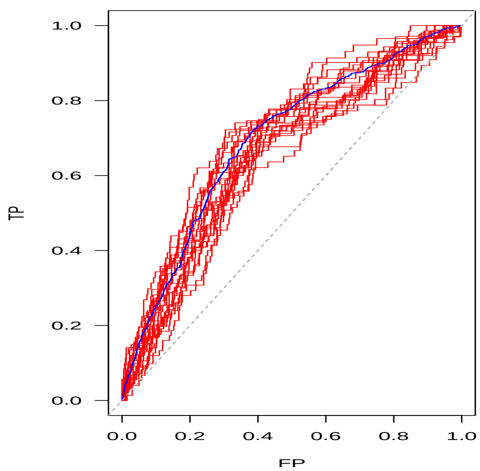

Fig. 5. ROC for $logit_2$ for weigthed estimation-jackknife.

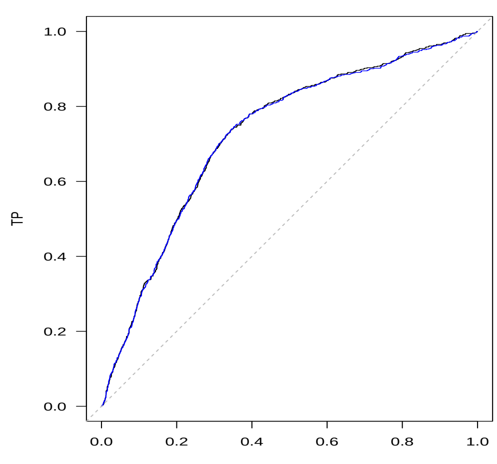

Fig. 6. ROC for $logit_3$ for unweighted (black curve) and weigthed (blue curve) estimation.

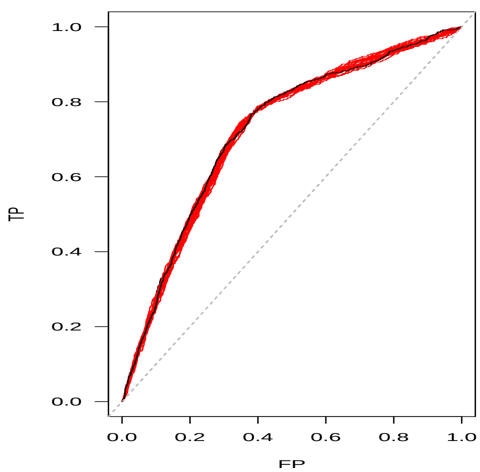

Fig. 7. ROC for $logit_3$ for unweighted estimation-bootstrap.

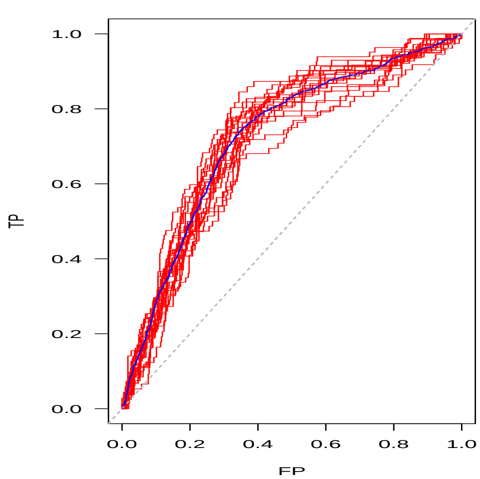

Fig. 8. ROC for $logit_3$ for weighted estimation-bootstrap.

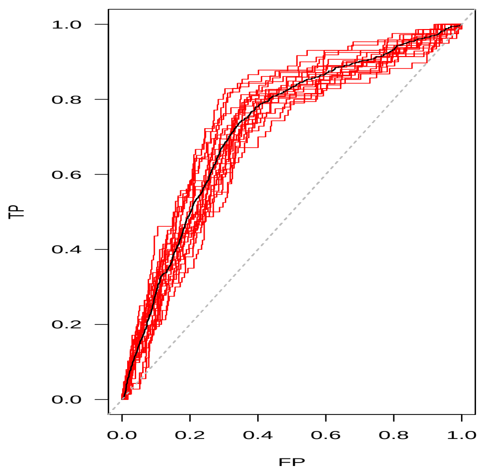

Fig. 9. ROC for $logit_3$ for unweighted estimation-jackknife.

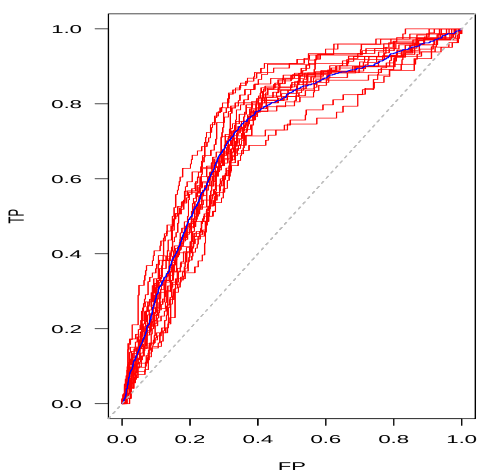

Fig. 10. ROC for $logit_3$ for weigthed estimation-jackknife.

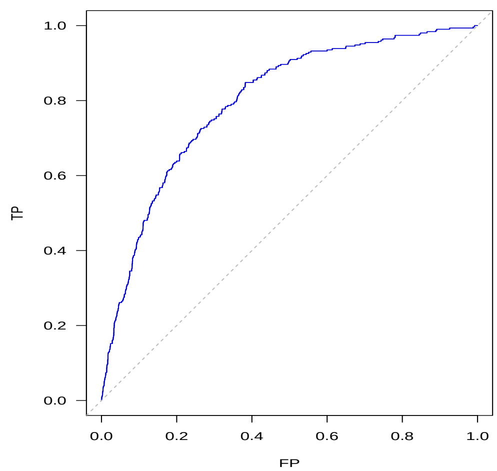

Fig. 11. ROC for $logit_4$ for unweighted (black curve) and weigthed (blue curve) estimation.

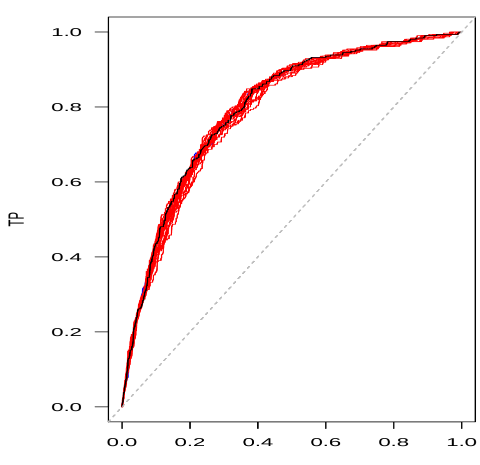

Fig. 12. ROC for $logit_4$ for unweighted estimation-bootstrap.

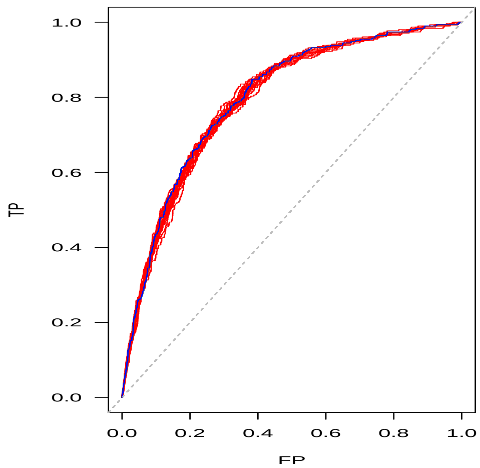

Fig. 13. ROC for $logit_4$ for weighted estimation-bootstrap.

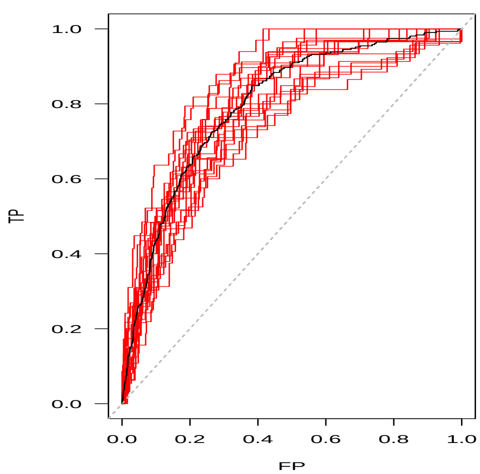

Fig. 14. ROC for $logit_4$ for unweighted estimation-jackknife.

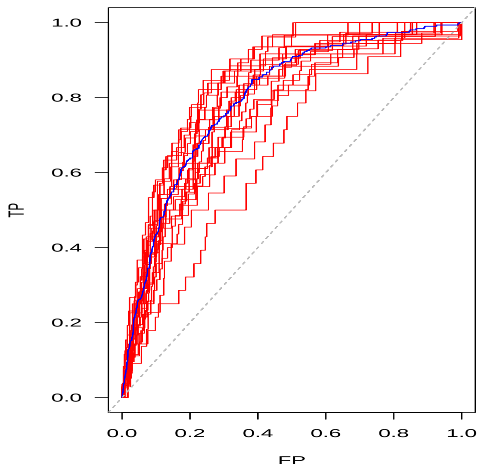

Fig. 15. ROC for $logit_4$ for weigthed estimation-jackknife.

References

1. Samoliński B, Raciborski F, Lipiec A i wsp.. (2014), Epidemiologia Chorób Alergicznych w Polsce (ECAP), in Polish *Alergol Pol 2014; 1: 10-8*
2. Tallo T. E. and A. Musdholifah A. (2018), The Implementation of Genetic Algorithm in Smote (Synthetic Minority Oversampling Technique) for Handling Imbalanced Dataset Problem, *2018 4th International Conference on Science and Technology (ICST), 2018, pp. 1-4,* doi: 10.1109/ICSTC.2018.8528591
3. King G and Zeng L, (2001), Logistic regression in rare events data, *Polit. Anal. 9 (2001), 137–163*
4. Zhang H., Li Z., Shahriar H., Tao L., Bhattacharya P., Qian Y, (2019), Improving Prediction Accuracy for Logistic Regression on Imbalanced Datasets, *IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), 2019, 918-919* doi: 10.1109/COMPSAC.2019.00140
5. Zhang L., Geisler T., Ray H., Xie Y (2021), Improving logistic regression on the imbalanced data by a novel penalized log-likelihood function, *Journal of Applied Statistics, 1-21*