Journal of KONBiN 43(2017)
ISSN 1895-8281
DOI 10.1515/jok-2017-0038
ESSN 2083-4608

THE POWER OF PREDICTION PROCESS IN THE TESTS OF AMMUNITION ELEMENTS

MOC PROCESU PREDYKCJI W BADANIACH ELEMENTÓW ŚRODKÓW BOJOWYCH

Dariusz Ampuła

Military Institute of Armament Technology
Wojskowy Instytut Techniczny Uzbrojenia

Abstract: *The author reminds the definition of coefficient determination and the idea of the largest credibility method in the introduction of the article. Firstly, the aspect of new coefficient defined by McFadden as a power of prediction process and called pseudo R^2 was characterized. The similar type of R^2 coefficients proposed by other known statisticians were described. Moreover, the hierarchic way of building the logistic regress model, through adding next variables to adjust an estimated model to empirical data was introduced. Four kinds of variables were analysed, whose adjustment influence on the estimated logistic regress model was affected by the quantity of inconsistences which appeared in the data result. The calculations of parameter R^2 made by the McFadden's, Nagelkerke's and Cox-Snell's formulas were presented. Concise conclusions relating to the estimated logistic regress model on the basis of empirical data from tested MD-8 fuses type were introduced in the end of the article and this model was compared to values of R^2 coefficient counted in the article. It was stated, that McFadden's pseudo R^2 parameter is most often used and it defines the power of prediction process.*

Keywords: *coefficient of determination (R-squared), logistic regress, credibility, variable, power of prediction process.*

Streszczenie: *W artykule scharakteryzowano postać nowego współczynnika określonego przez McFadden-a jako moc procesu predykcji i nazywanego pseudo R^2. Przedstawiono praktyczny przykład określania mocy predykcji w oparciu o wybrane dane empiryczne. Zaprezentowano hierarchiczny sposób budowy modelu regresji logistycznej, poprzez dołączanie kolejnych zmiennych w celu dopasowania szacowanego modelu do danych empirycznych. Analizowano cztery rodzaje zmiennych, których wpływ na dopasowanie szacowanego modelu regresji logistycznej był uzależniony od ilości niezgodności jakie wystąpiły w wynikach danych. Przedstawiono obliczenia parametru R^2 wykonane za pomocą wzorów McFadden-a, Nagelkerke-a i Cox-Snell-a. Na końcu artykułu przedstawiono zwięzłe wnioski dotyczące oszacowanego modelu regresji logistycznej na podstawie danych empirycznych z badanych zapalników typu MD-8 oraz porównano ten model do obliczonych w artykule wartości współczynnika R^2. Stwierdzono, że parametr pseudo R^2 McFadden'a jest najczęściej używany i określa on moc procesu predykcji.*

Słowa kluczowe: *współczynnik determinacji, regresja logistyczna, wiarygodność, zmienna, moc procesu predykcji,*

THE POWER OF PREDICTION PROCESS IN THE TESTS OF AMMUNITION ELEMENTS

1. Introduction

The way of prediction by means the method of logistic regress is currently more widely applied in different fields of knowledge. This regress is related to essential parameter which is the concept of the power of prediction process definition. This concept begins to be also used in other kinds of regress. The purpose of this article is to introduce to the reader different ways of designating of determined parameter as the power of prediction process and also show the practical way of his designating.

The parameter called coefficient of determination R^2 appears in the simple and multiple regress analysis, which is used to describe degree of foreseen independent variable adjustment. It is worth remembering that this parameter measures the part of the general variability of dependent variable is explained by the linear regress.

The issue of constructing equivalent measures for the logistic regress or the probit regress was an object of inquiries statisticians for many recent years. These measures in the majority base on values of credibility function of analysed model and the model with only absolute term. The method of greatest credibility of analysed model is widely described in the statistical literature. A fundamental idea of the greatest credibility method for evaluations of estimated parameters is to take values with the greatest credibility.

2. Evaluation of power prediction

The coefficient of determination R^2, as known is counted as quotient of explained by regress of the squares sum to total of the squares sum. In the other words it could be said that this coefficient also called as determinateness coefficient is a measure of percentage variability of dependent variable (explained) is explained by means of independent variable (the factor explanatory variable, predictor). The coefficient of determination informs how analysed model and tested factor explains gathered empirical data.

In the logistic regress analysis the role of total squares sum can be expressed by: $-2\ln L_0$, where $\ln L_0$ is a logarithm function of credibility for the model containing the absolute term only. However, the part of explained by regress of the squares sum fulfils expression:

$$(-2lnL_0\,) - (\,-2lnL_p\,) \tag{1}$$

where: lnL_p – logarithm of credibility function for full model.

According to [1] and [2] this observation inspired Daniel McFadden's to propose the definition of the new coefficient given formula in the 1974:

$$R^2_{MF} = 1 - \; lnL_p/lnL_0 \tag{2}$$

This coefficient appears in the literature from this time under name pseudo $R^{\,2}$ and by many statisticians it is determined as the power of carried out prediction process. This statistic can be interpreted as a proportional reduction of credibility logarithm. Additionally, it shows how better is the adjustment of empirical data in chosen model from naive model, in which the only explanatory variable is the constant. Therefore, in a manner of speaking it shows how variance of independent variables clearly explain the differentiation of dependent variable.
When a new independent variables is added to analysed model, this value can be easily calculate and grows. It can be also be proved that the greater number of observations in the attempt on the basis which are estimated parameters of the model, the lower value of McFadden's parameter. Additionally this coefficient never achieves the value 1.
The least pseudo R^2 value is equal to zero and appears in the situation when the proposed model does not bring anything new. This means, that all parameters besides the absolute term are insignificant that is do not differ from zero, i.e. expression (1) is equal zero. If the value pseudo R^2 is greater, this better is expectation of variable by means our determined model. The values close 1 show the best fitting of the model. Therefore, pseudo R^2 is very similar to R^2 applied in multiple regress analysis. A defect of this coefficient if fact, that his maximum value is smaller from the value 1.
The McFadden coefficient, copy the same unpleasant in uses property like standard coefficient of determination R^2. As already I remembered earlier, his value grows after adding the next explanatory variable to our model. From this regard we should use from formula on corrected value of pseudo R^2 McFadden's coefficient, which has form:

$$R^2_{MF} = 1 - (lnL_p - K)/lnL_0 \tag{3}$$

where: K – the number of variables estimated model.

The value of this coefficient is so corrected for the number of variables taken into account in the functional form of our analysed model.
Other statisticians also undertaken an attempt definition some parameter, which would be able to say about the degree adjustment of foreseen independent variable, determined by means analysed of prediction model. Somewhat earlier in 1970 Cragg and Uhler determined their parameter R^2 as:

$$R^2_{CU} = \; (1 - \left(L_0/L_p\right)^{2/n})/(1 - (L_0)^{2/n}) \tag{4}$$

where: n – the number of observations.

This coefficient accepts also values from the range [0, 1]. The value of this statistics, similar as value of McFadden's statistics depends on the vitalness statistical estimated model that is on total vitalness evaluations of unknown parameters.
The next statistician, who presented his value of R^2 coefficient was in 1991 Nico Nagelkerke. He tried modify value of McFadden's coefficient so, that his maximum value accepts the value 1. The value of R^2 parameter he determined as formula:

$$R_N^2 = \left(1 - e^{-(2/n)(lnL_p - lnL_0)}\right) / \left(1 - e^{(2/n)lnL_0}\right) \tag{5}$$

where: $e \approx 2{,}718$ – the Euler's number.

Sometimes in the literature and in [3] we can find definition of this R^2 Nagelkerke parameter appointed as $\bar{R}^2$.
Other value of this parameter determined Cox and Snell. According to them his value should be counted according to formula:

$$R_{CS}^2 = 1 - e^{-((-2lnL_0)-(-2lnL_p)/n)} \tag{6}$$

The next statistician, who proposed his form of R^2 parameter was G.S. Maddala. According to him [9] this statistics should have form:

$$R_M^2 = 1 - lnL_0/lnL_p \tag{7}$$

The construction of this statistics according to Maddala's assumes, that the random component has a normal distribution. It has close properties to McFadden's statistics. Her upper limit is also smaller from the value 1.
According to [4] and [5] exist still a lot of other R^2 coefficients appointed by such statisticians as: McKelvey and Zavoina (1975), Aldrich and Nelson (1984), Kvalseth (1985), Magee (1990), Mittlbock and Schemper (1996), Menard (2000), Tjur (2009), but generally in most cases pseudo R^2 McFadden's coefficient is most often appointed by statistical software and it is most often used by statisticians dealing with these problematic aspects.
As it is visible in presented dependences, the approach of statisticians to determine this parameter was somewhat differ. We also should be very careful while comparing measures origin from differ articles, unless we are sure, that this is the same R^2 coefficient.

3. Practical example defined power prediction

The practical example calculated value of pseudo R^2 coefficient by means the software [6] will be introduced bellow for analysed data results which were obtained during test of MD-8 fuse types. This type of fuses are base fuses and are applied in armour – piercing – tracer 76 mm, 85 mm and 100 mm cartridges. In this example we will determine model of logistic regress for these analysed fuses.

During the analysis of these test results, only laboratory test results were taken. Results obtained during test using special program and during shooting tests were not taken into account. Test results during so called scientifically – research inquiries were also eliminated. All these restrictions are aimed at obtaining uniform and clear test results reached with the same test procedures.

We will analyse the influence of chosen tested features elements of MD-8 fuses on taken postdiagnostic decisions and the same we will determine the hierarchical way of build predictive model of logistic regress for the purpose definition value of pseudo R^2 parameter.

The analysis consist following marks presented in the table 1. Variable Y is dependent variable describing accepted positive postdiagnostic decision „1" or negative postdiagnostic decision „0". Defining probability obtainment of positive postdiagnostic decision in the aim of tests, therefore its assigned value is 1. This variable was determined in the dichotomous form, therefore the only values 0 or 1. Positive postdiagnostic decisions (B5, B3) mean, that based on previous laboratory tests any negative influence on the further normal exploitation of tested fuses were not found. The figure 5 or 3 means the period of safe exploitation after which fuse lots should be diagnosed again. Negative postdiagnostic decisions consist all remaining accepted decisions.

Variables $X1 \div X3$ are qualitative variables which are results of individual analysed features tested. The zero value means, that inconsistency in the tested feature did not appear. However, the value 1 means, that this inconsistency appeared. Variable $X4$ is quantitative independent variable describing the age of tested fuses lot in years in the test moment.

Following tested features of MD-8 fuses were analysed: $X1$ means qualitative independent variable describing tested feature inertial spring of this fuse, $X2$ means qualitative independent variable describing appearing internal corrosion in fuses determined in the test methodology as classes from VI÷IX [10] and $X3$ means qualitative independent variable describing incompatible working fire element of fuse which is the primer cap.

In the table 1 test results for 224 test samples of MD-8 fuses, prepared to analysis in the dichotomous form for qualitative variables were presented. This table contains randomised information necessary to carry out definition parameter power of prediction process. Please note that in case of logistic regress analysis qualitative variables have to be introduced in the dichotomous form. It means that they should be suitably prepared for analysis of prediction process by means logistic regress.

Table 1 – Test results for test samples of MD-8 fuses

Lp.	Y	X1	X2	X3	X4	Lp.	Y	X1	X2	X3	X4	Lp.	Y	X1	X2	X3	X4
1	1	0	0	0	21	76	1	0	0	0	19	151	0	0	0	0	26
2	0	0	1	0	20	77	0	0	1	1	15	152	1	0	0	0	28
3	1	0	0	0	23	78	1	0	0	0	20	153	1	0	0	0	33
4	1	0	0	0	17	79	1	0	0	0	14	154	0	0	0	1	34
5	1	0	0	0	22	80	1	0	0	0	15	155	1	0	0	0	28
6	0	0	0	1	16	81	1	0	0	0	16	156	1	0	0	0	31
7	0	0	0	1	16	82	1	1	0	0	21	157	1	0	0	0	35
8	0	0	0	1	16	83	1	0	0	0	15	158	1	0	0	0	28
9	1	0	0	0	22	84	0	0	0	0	19	159	0	0	0	1	27
10	1	0	0	0	23	85	0	0	0	0	21	160	1	0	0	0	27
11	1	0	0	0	17	86	1	0	0	0	21	161	1	1	0	0	36
12	1	0	0	0	22	87	1	0	0	0	16	162	0	1	0	0	36
13	0	0	0	1	16	88	0	1	0	1	22	163	1	0	0	0	29
14	0	0	0	1	16	89	1	0	0	0	16	164	1	0	0	0	15
15	1	0	0	0	21	90	1	0	0	0	25	165	1	0	0	0	27
16	1	0	0	0	17	91	1	0	0	0	26	166	1	0	0	0	29
17	1	0	0	0	16	92	1	0	1	0	30	167	1	0	0	0	33
18	1	1	0	0	22	93	1	0	0	0	20	168	1	1	0	0	39
19	1	0	0	0	20	94	1	0	0	1	22	169	1	0	0	0	37
20	1	0	0	0	24	95	1	0	0	0	22	170	1	0	0	0	29
21	1	0	0	0	17	96	1	0	0	0	25	171	1	0	0	0	29
22	1	0	0	0	18	97	1	0	0	0	16	172	1	1	0	0	39
23	0	0	0	1	19	98	1	0	0	0	17	173	0	1	0	0	37
24	0	0	0	1	20	99	1	0	0	0	16	174	0	1	0	0	32
25	0	0	1	0	25	100	0	0	0	1	18	175	1	0	0	0	30
26	0	0	1	1	28	101	1	0	0	0	19	176	1	0	0	0	30
27	1	0	0	0	25	102	0	0	0	1	33	177	1	0	0	0	31
28	0	0	0	1	21	103	1	1	0	0	27	178	0	0	0	0	37
29	0	0	1	0	23	104	0	1	0	1	26	179	0	1	0	0	39
30	1	0	0	0	22	105	0	1	0	0	26	180	0	1	0	0	39
31	0	0	1	0	30	106	1	0	0	0	27	181	1	0	0	0	33
32	1	0	0	0	23	107	1	0	0	0	28	182	1	0	0	0	39
33	0	0	0	0	24	108	1	0	0	0	29	183	0	1	0	0	41
34	0	0	1	1	26	109	1	0	0	0	28	184	0	0	0	0	40
35	0	0	1	1	27	110	1	0	0	0	22	185	0	0	0	0	32
36	0	0	1	0	24	111	1	0	0	0	21	186	1	0	0	0	37
37	0	0	1	0	30	112	1	0	0	0	23	187	0	1	0	0	41
38	0	0	0	1	26	113	1	0	0	0	22	188	0	1	0	0	41
39	0	0	1	0	30	114	1	0	0	0	22	189	0	0	0	0	34
40	0	0	1	1	26	115	1	0	0	0	26	190	1	1	0	0	42

Table 1(continued) – Test results for test samples of MD-8 fuses

41	0	0	1	1	27	116	1	1	0	0	30	191	0	1	0	0	41
42	0	0	0	1	26	117	1	0	0	0	31	192	1	0	0	0	35
43	0	0	0	1	26	118	1	0	0	0	22	193	0	0	0	0	35
44	0	0	1	1	26	119	1	0	0	0	10	194	1	0	0	0	33
45	0	0	0	1	26	120	0	0	0	1	31	195	1	0	0	0	52
46	0	0	0	1	26	121	1	0	0	0	22	196	1	0	0	0	52
47	0	0	0	1	26	122	1	0	0	0	21	197	0	0	0	1	52
48	0	0	0	1	27	123	1	0	0	0	23	198	0	1	0	0	42
49	0	0	0	1	27	124	1	0	0	0	23	199	1	0	0	0	34
50	1	0	0	0	27	125	1	0	0	0	32	200	1	0	0	0	34
51	0	0	0	1	26	126	0	0	0	1	24	201	1	0	0	0	11
52	1	0	0	0	10	127	1	0	0	0	25	202	0	0	0	1	11
53	1	0	0	0	11	128	1	0	0	0	24	203	0	1	0	0	44
54	1	0	0	0	11	129	0	1	0	0	31	204	1	0	0	0	46
55	1	0	0	0	15	130	0	0	0	1	28	205	1	1	0	0	37
56	0	0	0	1	16	131	0	0	0	0	28	206	1	1	0	0	37
57	1	0	0	0	10	132	0	0	0	0	28	207	1	0	0	0	38
58	1	0	0	0	11	133	0	1	0	1	28	208	1	0	0	0	37
59	1	0	0	0	16	134	0	0	0	0	28	209	1	0	0	0	37
60	1	0	0	0	10	135	0	0	0	0	22	210	0	1	0	0	44
61	1	0	0	0	15	136	0	0	0	0	31	211	1	0	0	0	40
62	1	0	0	0	14	137	0	0	0	0	24	212	1	0	0	0	47
63	1	0	0	0	15	138	0	0	0	1	33	213	0	1	0	0	46
64	1	0	0	0	15	139	0	0	1	1	43	214	0	0	0	0	38
65	0	0	0	1	15	140	1	0	0	0	25	215	1	0	0	0	40
66	1	0	0	0	10	141	0	1	1	0	32	216	1	0	0	0	41
67	0	0	0	0	10	142	1	0	0	0	33	217	0	0	0	0	41
68	0	0	0	0	16	143	0	0	0	0	33	218	0	0	0	0	40
69	1	0	0	0	13	144	0	1	0	0	34	219	0	1	0	0	49
70	1	0	0	0	15	145	1	1	0	0	35	220	1	0	0	0	42
71	1	0	0	0	10	146	1	1	0	0	34	221	1	0	0	0	42
72	1	0	0	0	16	147	1	0	0	0	26	222	0	1	0	0	50
73	1	0	0	0	10	148	1	0	0	0	34	223	1	0	0	0	46
74	1	0	0	0	12	149	1	0	0	0	34	224	1	0	0	0	46
75	1	0	0	0	18	150	1	0	0	0	26						

As the first independent variable to the model, *X1* variable will be introduced. Thanks to the software usage [6] results sheet visible in figure 1 will be received.

Model: logistic regression (logit) N zeros: 87 ones: 137 (MD-8)
Dependent variable: Y Loss: Greatest prob. average square error scal
Total loss: 144,13462807 Chi2(1)=11,005 p=,00091

N=224	Constant B0	X1
Evaluation	0,6539265	-1,260062
Standard error	0,1529211	0,3900999
t(222)	4,276235	-3,230102
p	0,00002824008	0,00142516
-95%CL	0,3525637	-2,028835
+95%CL	0,9552892	-0,4912896
Chi-square Wald's	18,28618	10,43356
p	0,00001906313	0,001238711
Quotient of chances with individ	1,923077	0,2836364
-95%CL	1,42271	0,1314886
+95%CL	2,599422	0,6118369
Quotient of chances with range		0,2836364
-95%CL		0,1314886
+95%CL		0,6118369

Fig. 1 Sheet of results for $X1$ variable

As it is visible from result sheet presented, $X1$ variable and value of $B0$ constant are essential values. The calculated model can be well-adjusted to analysed data. The received logistic model would have form (only in case of independent variable $X1$):

$$P\ (Y = 1) = e^{0{,}6539265 - 1{,}26006X1} / 1 + e^{0{,}6539265 - 1{,}26006X1} \tag{8}$$

Within so accepted marks ($Y = 1$ means accepted positive decision), it is visible in the sheet quotient of chances for the individual change of $X1$ variable equals *0,28* informs that probability obtainment of positive postdiagnostic decision at the occurrence of this type of inconsistencies is *0,28* times greater than at her nonoccurrence.

Wanting to check the next variables taken out of model to analysis, we build further hierarchical logistic model, so we join next variable – values of $X2$ independent variable. We will then receive the following results sheet of these variables visible in figure 2.

In case of addition of $X2$ variable we see that all variables are also statistical essential. We can find, that model adjustment of analysed empirical data gives new statistical essential logistic model.

The received logistic model has form:

$$P\ (Y=1)=\frac{e^{0{,}908676-1{,}470238X1-3{,}632685X2}}{1+e^{0{,}908676-1{,}470238X1-3{,}632685X2}} \tag{9}$$

	Model: logistic regression (logit) N zeros: 87 ones: 137 (MD-8) Dependent variable: Y Loss: Greatest prob.average square error scal Total loss: 129,74622862 Chi2(2)=39,782 p=,00000		
N=224	Constant B0	X1	X2
Evaluation	0,9086756	-1,470238	-3,632685
Standard error	0,1675084	0,3984131	1,04511
t(221)	5,424657	-3,690234	-3,475888
p	0,0000001519663	0,0002821941	0,000612812
-95%CL	0,5785574	-2,255413	-5,692342
+95%CL	1,238794	-0,6850625	-1,573029
Chi-square Wald's	29,42691	13,61783	12,0818
p	0,00000005848621	0,000224419	0,0005098385
Quotient of chances with individ	2,481035	0,2298709	0,02644508
-95%CL	1,783464	0,1048303	0,003371687
+95%CL	3,451448	0,5040587	0,2074161
Quotient of chances with range		0,2298709	0,02644508
-95%CL		0,1048303	0,003371687
+95%CL		0,5040587	0,2074161

Fig. 2 Sheet of results for $X1$ and $X2$ variables

The next variable added to the model will be $X3$ independent variable. In case so calculated model for three analysed variables, this model is also statistically essential. As it is visible from result sheets presented in figure 3, all variables are important i.e. $B0$ constant and $X1, X2$ *and* $X3$ variables are statistically essential values. Moreover, analysed $X3$ variable added to model does not influence on change or damage of model adjustment to analysed empirical data. The received logistic model at the consider three variables will have form (10).

$$P\ (Y=1)=\frac{e^{1{,}769-2{,}179X1-3{,}741X2-5{,}155X3}}{1+e^{1{,}769-2{,}179X1-3{,}741X2-5{,}155X3}} \tag{10}$$

Let's see now, how our model will look with the last $X4$ variable. As we know this variable tells the age of tested lots of MD-8 type fuses. The $B0$ constant and $X1$, $X2$, $X3$ *and* $X4$ variables are statistically essential. Joining this variable to the model did not have negative influence to adjustment of the model to analysed empirical data.

The figure 4 in result sheets for all variables which were analysed in adjustment statistical model are presented as a proof.

	Model: logistic regression (logit) N zeros: 87 ones: 137 (MD-8) Dependent variable: Y Loss: Greatest prob.average square error scal. Total loss: 87,442386148 Chi2(3)=124,39 p=0,0000			
N=224	**Constant B0**	X1	X2	X3
Evaluation	1,769525	-2,178737	-3,740751	-5,155468
Standard error	0,2360653	0,440248	1,090627	1,04277
t(220)	7,49591	-4,948886	-3,42991	-4,944011
p	1,59409100E-12	0,000001483405	0,0007208831	0,000001517187
-95%CL	1,304286	-3,04638	-5,890164	-7,210566
+95%CL	2,234764	-1,311094	-1,591338	-3,100371
Chi-square Wald's	56,18867	24,49147	11,76428	24,44325
p	6,75407800E-14	0,0000007501813	0,0006045433	0,0000007691801
Quotient of chances with individ.	5,868063	0,1131844	0,02373627	0,005767778
-95%CL	3,685056	0,04753068	0,002766523	0,0007387388
+95%CL	9,344273	0,2695251	0,203653	0,0450325
Quotient of chances with range		0,1131844	0,02373627	0,005767778
-95%CL		0,04753068	0,002766523	0,0007387388
+95%CL		0,2695251	0,203653	0,0450325

Fig. 3 Sheet of results for $X1, X2$ *and* $X3$ *variables*

	Model: logistic regression (logit) N zeros: 87 ones: 137 (MD-8) Dependent variable: Y Loss: Greatest prob. average square error scal. Total loss: 84,156366921 Chi2(4)=130,96 p=0,0000				
N=224	**Constant B0**	X1	X2	X3	X4
Evaluation	3,272861	-1,64573	-3,797206	-5,413678	-0,0553106
Standard error	0,6770578	0,4787516	1,095729	1,058868	0,02186745
t(219)	4,833946	-3,437544	-3,465461	-5,112703	-2,529358
p	2,51848200E-6	0,000702431	0,000636816	6,9176730E-7	0,01213089
-95%CL	1,938478	-2,58928	-5,95673	-7,500554	-0,0984082
+95%CL	4,607244	-0,7021797	-1,637683	-3,326802	-0,012213
Chi-square Wald's	23,36703	11,81671	12,00942	26,13974	6,397651
p	1,34473300E-6	0,000587761	0,000530017	3,1940780E-7	0,01143191
Quotient of chances with individ	26,38672	0,1928717	0,02243335	0,004455222	0,9461912
-95%CL	6,948166	0,07507406	0,002588362	0,000552778	0,9062789
+95%CL	100,2076	0,4955041	0,19443	0,03590773	0,9878613
Quotient of chances with range		0,1928717	0,02243335	0,004455222	0,09797471
-95%CL		0,07507406	0,002588362	0,000552778	0,01603238
+95%CL		0,4955041	0,19443	0,03590773	0,5987285

Fig. 4 Sheet of results for $X1, X2, X3$ *and* $X4$ *variables*

The received logistic model with four variables have form:

$$P\ (Y = 1) = \frac{e^{3,27-1,65X1-3,8X2-5,41X3-0,06X4}}{e^{3,27-1,65X1-3,8X2-5,41X3-0,06X4}} \tag{11}$$

In figures 1 ÷ 4 in result sheets for analysed variables appointed as Chi-square Wald's we have values of test or else Wald's statistics which checks a vitalness of parameters and the line below there are values of probability level p connected with it that is the kindness adjustment level. While building and enlarging our model for one variable as in our case, it is more practical to use Wald's test. We receive it from counting the square of quotient estimated coefficient of regress through its standard error. As it is visible from calculating values of this test where these values somewhat differ. However, the probability levels are on highly essential level and we can therefore draw a conclusion, that all analysed variables have essential influence on taken postdiagnostic decisions.
The probability level p for $t - Student's$ test assisting to essential tests coefficients of regress is for all variables on $p < 0{,}1$ level that is lower than accepted essential level. It is on the highly essential level and this means, that regress parameters are also statistically essential.
After the defining our model to analysed variables, let's count value of R^2 coefficient for this case. We will compare calculated values power of prediction process for four proposed models presented in the article, from each is statistically essential. The results of R^2 parameter value calculated by means of forms (2), (3), (5) and (6) and thanks to [6] software, which counts values of logarithms are presented in table 2.

Table 2 – Results of calculations R^2 values

The variables taken into account in analysed model	R^2 McFadden's irrespective number of variables	R^2 McFadden's with taken into account number of variables	R^2 Nagelkerke	R^2 Cox-Snell's
1	2	3	4	5
X1	0,036774	0,040115	0,065042	0,047944
X1 i X2	0,132929	0,139612	0,220754	0,162722
X1, X2 i X3	0,415638	0,425662	0,578072	0,426107
X1, X2, X3 i X4	0,437598	0,450964	0,600583	0,442700

The short analysis calculation value of R^2 parameter testifies about analysed models are adjusted to empirical data. However, the discriminant in the definition power adjustment of analysed model or power of prediction process is counted R^2 parameter. Moreover, looking at values of this parameter we can find, that the best adjusted model is with all analysed variables, because value of R^2 parameter is the highest.

However, we should not lean to this ratio, because as mentioned earlier a defect of this situation is fact, that increasing independent variables in the model comes with a grow of R^2 parameter value. Therefore, the most statisticians consider McFadden's coefficient as most credible R^2 parameter with taken into account quantity of analysed variables, despite the fact, that coefficient or Nagelkerke parameter obtained greater values. Closer to McFadden's coefficient are calculated values of R^2 Cox-Snell's parameter. In the table 2 shows values of pseudo R^2 McFadden's parameter with taken into account number of variables, are contained in the column 3.

4. References

The conducted example analysis was an attempt of adjustment model of logistic regress to empirical data obtained during test of base MD-8 fuse type for the purpose definition of pseudo R^2 coefficient that is the purpose of defining power of prediction process. The test results of these fuses were received during their laboratory tests. This analysis confirms the opinion of many statisticians, that R^2 coefficient has some defects. They are mainly: the sensibility on the variables number applied in the model and lack of the possibility to achieve the upper limit equal the 1.
In the statistical literature we can find definition, that value of pseudo R^2 McFadden's parameter contained between values $0{,}2 \div 0{,}4$, testifies about the proper adjustment of model. On this base we can find, that the adjustment our analysed model can find as satisfying. In our presented case value of this coefficient for four analysed variables carried out $pseudo\ R^2 = 0{,}450964$, that is a greater value from recommendations of many statisticians. The power of prediction process is so large what means, that analysed model with four variables were well-adjusted.
The way of analysing of received empirical data presented in the article shows, that in many cases during model adjustment, all chosen variables can have essential influence on searching adjustment of logistic model to analysed empirical data. These variables except this, that are essential have also huge influence on definition of correct prediction of safe exploitation period length of tested elements of ammunition.
Summing up, we can find, that despite the fact, that calculated values of R^2 Nagelkerke coefficient (table 2) have higher values, this does not mean, that power of prediction process is on this level for analysed adjustment model. Prominent statisticians continuously try to carry out a lot of proofs and analyses in the theme the choice of the most credible R^2 coefficient. As I already mentioned, many of them after thorough tests and lots of carried scientific inquiries out think that pseudo R^2 McFadden coefficient is this the most credible measure.
It is also an important aspect at the adjustment of model to analysed empirical data is to have good knowledge of specialist statistical computer software which usage now is indispensable to reach and obtain correct value of searching parameter power of prediction process.

5. Literature

[1] A. Stanisz – *The accessible course of the statistics* – Statsoft Poland, Kraków 2007,
[2] *Wikipedia* – the free encyclopaedia,
[3] *Naukowiec.org*. – the internet portal,
[4] P. Allison – *What's the Best R-Squared for Logistic Regression?* – Statistical Horizons, 2013,
[5] M. Veall, K.. Zimmermann – *Pseudo-R2 Measures for Some Common Limited Dependent Variable Models* – 1996,
[6] Statistica 12 – *Statsoft Poland 2012* – computer software,
[7] M. Rabiej – *The statistics with the program Statistica* – the publishing house Helion, Gliwice 2012,
[8] S. Kot, J, Jakubowski, A. Sokołowski – *The statistics* – the publishing house Difin, Warsaw 2011,
[9] G. S. Maddala – *The econometrics* – the publishing house PWN, Warsaw 2008,
[10] The group work – *The methodology of diagnostic tests of ammunition* – Index N-5001b – 1985, Military Institute of Armament Technology.

D.Sc. Eng. Dariusz Ampuła, *graduated Military Technical Academy in the area of the Ground Armament Systems. He became the scientifically – test worker at Military Institute of Armament Technology in Zielonka after holding of one year professional practice. The degree scientific of doctor he obtained in 2006 in Air Force Institute of Technology. The author and co–author of works from the area technical diagnostics and reliability of working land ammunition. In his analytical analysis he focuses on the problems of prediction process tested elements of land ammunition.*

MOC PROCESU PREDYKCJI W BADANIACH ELEMENTÓW ŚRODKÓW BOJOWYCH

1. Wstęp

Obecnie coraz szerzej stosowany jest w różnych dziedzinach wiedzy sposób predykcji za pomocą metody regresji logistycznej. Z regresją tą wiąże się dość istotny parametr jakim jest pojęcie określenia mocy procesu predykcji. Pojęcie to zaczyna być także używane przy innych rodzajach regresji. Celem tego artykułu jest zapoznanie czytelnika z różnymi sposobami wyznaczania parametru określonego jako moc procesu predykcji, a także pokazanie praktycznego sposobu jego wyznaczania.
W analizie regresji prostej i wielorakiej występuje parametr o nazwie współczynnik determinacji R^2, który służy do określenia stopnia dopasowania przewidywanej zmiennej niezależnej. Przypomnę tylko, że parametr ten mierzy nam, jaka część ogólnej zmienności zmiennej zależnej jest wyjaśniona przez regresję liniową.
Zagadnienie skonstruowania równoważnych miar dla regresji logistycznej (logitowej) lub regresji probitowej było przedmiotem dociekań statystyków od wielu lat. Miary te w większości bazują na wartościach funkcji wiarygodności analizowanego modelu i modelu tylko z wyrazem wolnym. Metoda największej wiarygodności analizowanego modelu jest szeroko opisywana w literaturze statystycznej. Zasadniczą ideą metody największej wiarygodności jest to, aby za oceny szacowanych parametrów brać te wartości, dla których wiarygodność jest największa.

2. Ocena mocy predykcji

Jak wiadomo, współczynnik determinacji R^2 jest obliczany jako iloraz wyjaśnionej przez regresję sumy kwadratów do całkowitej sumy kwadratów. Innymi słowy można powiedzieć, że ten współczynnik zwany także współczynnikiem określoności jest miarą tego, jaki procent zmienności zmiennej zależnej (objaśnianej) jest wyjaśniany za pomocą zmiennej niezależnej (czynnik zmienna objaśniająca, predyktor). Czyli współczynnik determinacji informuje nas, ile nasz analizowany model, nasz badany czynnik wyjaśnia zgromadzone dane empiryczne.
W analizie regresji logistycznej rolę całkowitej sumy kwadratów może spełniać wyrażenie: $-2\ln L_0$, gdzie $\ln L_0$ jest logarytmem funkcji wiarygodności dla modelu zawierającego jedynie wyraz wolny. Natomiast rolę wyjaśnionej przez regresję sumy kwadratów spełnia wyrażenie:

$$(-2\mathrm{lnL_0}) - (-2\mathrm{lnL_p}) \qquad (1)$$

gdzie: lnL_p – logarytm funkcji wiarygodności dla pełnego modelu.

Według [1] oraz [2] to spostrzeżenie zainspirowało Daniela McFaddena do zaproponowania w roku 1974 określenia nowego współczynnika danego wzorem:

$$R^2_{MF} = 1 - \ln L_p / \ln L_0 \qquad (2)$$

Współczynnik ten występuje od tego czasu w literaturze pod nazwą *pseudo* R^2 i przez wielu statystyków jest określany jako moc przeprowadzonego procesu predykcji. Statystykę tę można interpretować jako proporcjonalną redukcję logarytmu wiarygodności. Można także powiedzieć, że pokazuje ona o ile lepiej dopasowany jest do danych empirycznych nasz wybrany model od modelu naiwnego, w którym jedyną zmienną objaśniającą jest stała. Zatem, w pewnym sensie pokazuje w jaki sposób wariancja zmiennych niezależnych pozwala wytłumaczyć zróżnicowanie zmiennej zależnej.
Wartość ta jest łatwa do wyliczenia i rośnie, kiedy dodajemy nowe zmienne niezależne do analizowanego modelu. Można także wykazać, że im większa jest liczba obserwacji w próbie na podstawie której szacowane są parametry modelu, tym niższą wartość przyjmuje ten parametr McFaddena. Dodatkowo ten współczynnik nigdy nie osiąga wartości jeden.
Wartość najmniejsza *pseudo* R^2 jest równa zero i występuje w sytuacji kiedy zaproponowany przez nas model nie wnosi nic nowego. Oznacza to, że wszystkie parametry poza wyrazem wolnym są nieistotne czyli nie różnią się od zera, tzn. wyrażenie (1) jest równe zero. Im wartość *pseudo* R^2 jest większe, tym lepsze jest przewidywanie zmiennej za pomocą naszego określonego modelu. Wartości bliskie jeden wskazują na idealne dopasowanie modelu. Zatem *pseudo* R^2 jest bardzo podobny do R^2 stosowanego w analizie regresji wielorakiej. Wadą tego współczynnika jest fakt, że jego wartość maksymalna jest mniejsza od wartości jeden.
Współczynnik McFadden-a, powiela tę samą nieprzyjemną w zastosowaniach własność co standardowy współczynnik determinacji R^2. Jak już wspomniałem wcześniej, jego wartość rośnie po dodaniu kolejnej zmiennej objaśniającej do naszego modelu. Z tego też względu powinno korzystać się ze wzoru na skorygowaną wartość współczynnika *pseudo* R^2 McFadden-a, który ma postać:

$$R^2_{MF} = 1 - (\ln L_p - K) / \ln L_0 \qquad (3)$$

gdzie: K – liczba zmiennych szacowanego modelu.

Wartość tego współczynnika jest korygowana więc o liczbę zmiennych uwzględnionych w postaci funkcyjnej naszego analizowanego modelu.
Inni statystycy także podjęli próbę określenia jakiegoś parametru, który mógłby powiedzieć o stopniu dopasowania przewidywanej zmiennej niezależnej, określonej za pomocą analizowanego modelu predykcyjnego.

Nieco wcześniej bo w roku 1970 Cragg i Uhler określili swój parametr R^2 jako:

$$R_{CU}^2 = \left(1 - \left(L_0/L_p\right)^{2/n}\right)/\left(1 - \left(L_0\right)^{2/n}\right) \tag{4}$$

gdzie: n – liczba obserwacji.

Współczynnik ten przyjmuje także wartości z przedziału [0, 1]. Wartość tej statystyki, podobnie jak wartość statystyki McFadden-a zależy od statystycznej istotności oszacowanego modelu czyli od łącznej istotności ocen nieznanych parametrów.
Kolejnym statystykiem, który przedstawił swoją wartość współczynnika R^2 był w roku 1991 Nico Nagelkerke. Próbował on modyfikować wartość współczynnika McFadden-a tak, aby jego wartość maksymalna przyjmowała wartość jeden. Wartość parametru R^2 określił jako wyrażenie:

$$R_N^2 = \left(1 - e^{-(2/n)(\ln L_p - \ln L_0)}\right)/\left(1 - e^{(2/n)\ln L_0}\right) \tag{5}$$

gdzie: $e \approx 2{,}718$ – liczba Eulera.

Czasami w literaturze oraz w [3] można spotkać określenie tego parametru R^2 Nagelkerke oznaczone jako $\bar{R}^2$.
Jeszcze inną wartość tero parametru określił Cox i Snell. Według nich wartość jego powinna być obliczana według wyrażenia:

$$R_{CS}^2 = 1 - e^{-((-2\ln L_0)-(-2\ln L_p)/n)} \tag{6}$$

Kolejnym statystykiem, który zaproponował swoją postać parametru R^2 był G.S. Maddala. Według niego [9] statystyka ta powinna mieć postać:

$$R_M^2 = 1 - \ln L_0/\ln L_p \tag{7}$$

Konstrukcja tej statystyki według Maddali zakłada, że składnik losowy ma rozkład normalny. Posiada ona zbliżone własności do statystyki McFadden-a. Jej górna granica jest także mniejsza od wartości jeden.
Według [4] i [5] istnieje jeszcze szereg innych współczynników R^2 wyznaczonych przez takich statystyków jak: McKelvey i Zavoina (1975), Aldrich i Nelson (1984), Kvalseth (1985), Magee (1990), Mittlbock i Schemper (1996), Menard (2000), Tjur (2009), ale ogólnie w większości przypadków współczynnik *pseudo* R^2 McFadden-a jest najczęściej wyznaczany przez oprogramowania statystyczne oraz jest on najczęściej używany przez statystyków zajmujących się tą problematyką.
Jak widać z przedstawionych zależności, podejście statystyków do określenia tego parametru było nieco odmienne. Należy więc być bardzo ostrożnym porównując miary pochodzące z różnych artykułów, o ile nie jesteśmy pewni, że chodzi o ten sam współczynnik R^2.

3. Praktyczny przykład określenia mocy predykcji

Poniżej przedstawiony zostanie praktyczny przykład obliczonej wartości współczynnika *pseudo* R^2 za pomocą oprogramowania [6] dla analizowanych wyników danych jakie zostały uzyskane podczas badania zapalników typu MD-8. Zapalniki tego typu są zapalnikami dennymi i stosowane są w nabojach przeciwpancerno – smugowych kalibru 76 mm, 85 mm oraz 100 mm. W przykładzie tym określać będziemy model regresji logistycznej dla tych analizowanych zapalników.

Podczas analizowania tych wyników badań, zostały wzięte pod uwagę tylko wyniki z badań laboratoryjnych. Nie uwzględniono wyników badań uzyskanych podczas badań według programu specjalnego oraz podczas badań strzelaniem. Wyeliminowano także wyniki badań podczas tzw. dociekań naukowo – badawczych. Wszystkie te ograniczenia miały na celu doprowadzenie do uzyskania jednolitych oraz jasnych wyników badań, uzyskanych według tych samych procedur badawczych.

Analizować będziemy wpływ wybranych badanych cech elementów zapalników MD-8 na podjęte decyzje podiagnostyczne a tym samym określać będziemy hierarchiczny sposób budowy predykcyjnego modelu regresji logistycznej w celu określenia wartości parametru *pseudo* R^2.

Do analizy przyjęto następujące oznaczenia przedstawione w tabeli 1. Zmienna Y jest zmienną zależną opisującą przyjętą dodatnią decyzję podiagnostyczną „1" lub ujemną decyzję podiagnostyczną „0". Interesuje nas określenie prawdopodobieństwa uzyskania dodatniej decyzji podiagnostycznej dlatego przypisano jej wartość jeden. Zmienna ta została określona w postaci dychotomicznej, czyli przyjmuje tylko wartości zero lub jeden. Dodatnie decyzje podiagnostyczne (B5, B3) oznaczają, że na podstawie przeprowadzonych dotychczasowych badań laboratoryjnych nie stwierdzono niezgodności mogących mieć wpływ na dalszą normalną eksploatację badanych zapalników. Cyfra 5 lub 3 oznacza okres bezpiecznej eksploatacji po którym partie zapalników powinny być ponownie zdiagnozowane. Do ujemnych decyzji podiagnostycznych zaliczono wszystkie pozostałe możliwe przyjęte decyzje.

Zmienne $X1 \div X3$ oznaczają zmienne jakościowe czyli wyniki poszczególnych analizowanych badanych cech. Wartość zero oznacza, że niezgodność w badanej cesze nie wystąpiła, natomiast wartość jeden oznacza, że niezgodność ta wystąpiła. Zmienna $X4$ jest zmienną niezależną ilościową opisującą wiek badanej partii zapalników w latach w chwili badania.

Analizowane były następujące badane cechy zapalników MD-8: $X1$ oznacza zmienną niezależną jakościową opisującą badaną cechę sprężyny bezwładnika tego zapalnika, $X2$ oznacza zmienną niezależną jakościową opisującą występującą korozję wewnętrzną w zapalnikach określoną w metodyce badawczej jako klasy od VI ÷ IX oraz $X3$ oznacza zmienną niezależną jakościową opisującą niezgodne działanie elementu ogniowego zapalnika jakim jest spłonka zapalająca.

Wyniki badań dla 224 próbek badawczych zapalników MD-8, przygotowane do analizy w postaci dychotomicznej dla zmiennych jakościowych, przedstawione zostały w tabeli 1. Tabela ta zawiera wybrane losowo informacje badawcze niezbędne do przeprowadzenia analizy w celu określenia parametru mocy procesu predykcji. Należy tutaj zwrócić uwagę, że w przypadku analizowania regresji logistycznej, zmienne jakościowe muszą być przedstawione w postaci dychotomicznej, czyli należy je odpowiednio przygotować do analizy procesu predykcji za pomocą regresji logistycznej.

Tabela 1 – Wyniki badań dla próbek badawczych zapalników MD-8

Lp.	Y	X1	X2	X3	X4	Lp.	Y	X1	X2	X3	X4	Lp.	Y	X1	X2	X3	X4
1	1	0	0	0	21	76	1	0	0	0	19	151	0	0	0	0	26
2	0	0	1	0	20	77	0	0	1	1	15	152	1	0	0	0	28
3	1	0	0	0	23	78	1	0	0	0	20	153	1	0	0	0	33
4	1	0	0	0	17	79	1	0	0	0	14	154	0	0	0	1	34
5	1	0	0	0	22	80	1	0	0	0	15	155	1	0	0	0	28
6	0	0	0	1	16	81	1	0	0	0	16	156	1	0	0	0	31
7	0	0	0	1	16	82	1	1	0	0	21	157	1	0	0	0	35
8	0	0	0	1	16	83	1	0	0	0	15	158	1	0	0	0	28
9	1	0	0	0	22	84	0	0	0	0	19	159	0	0	0	1	27
10	1	0	0	0	23	85	0	0	0	0	21	160	1	0	0	0	27
11	1	0	0	0	17	86	1	0	0	0	21	161	1	1	0	0	36
12	1	0	0	0	22	87	1	0	0	0	16	162	0	1	0	0	36
13	0	0	0	1	16	88	0	1	0	1	22	163	1	0	0	0	29
14	0	0	0	1	16	89	1	0	0	0	16	164	1	0	0	0	15
15	1	0	0	0	21	90	1	0	0	0	25	165	1	0	0	0	27
16	1	0	0	0	17	91	1	0	0	0	26	166	1	0	0	0	29
17	1	0	0	0	16	92	1	0	1	0	30	167	1	0	0	0	33
18	1	1	0	0	22	93	1	0	0	0	20	168	1	1	0	0	39
19	1	0	0	0	20	94	1	0	0	1	22	169	1	0	0	0	37
20	1	0	0	0	24	95	1	0	0	0	22	170	1	0	0	0	29
21	1	0	0	0	17	96	1	0	0	0	25	171	1	0	0	0	29
22	1	0	0	0	18	97	1	0	0	0	16	172	1	1	0	0	39
23	0	0	0	1	19	98	1	0	0	0	17	173	0	1	0	0	37
24	0	0	0	1	20	99	1	0	0	0	16	174	0	1	0	0	32
25	0	0	1	0	25	100	0	0	0	1	18	175	1	0	0	0	30
26	0	0	1	1	28	101	1	0	0	0	19	176	1	0	0	0	30
27	1	0	0	0	25	102	0	0	0	1	33	177	1	0	0	0	31
28	0	0	0	1	21	103	1	1	0	0	27	178	0	0	0	0	37
29	0	0	1	0	23	104	0	1	0	1	26	179	0	1	0	0	39
30	1	0	0	0	22	105	0	1	0	0	26	180	0	1	0	0	39
31	0	0	1	0	30	106	1	0	0	0	27	181	1	0	0	0	33
32	1	0	0	0	23	107	1	0	0	0	28	182	1	0	0	0	39
33	0	0	0	0	24	108	1	0	0	0	29	183	0	1	0	0	41
34	0	0	1	1	26	109	1	0	0	0	28	184	0	0	0	0	40
35	0	0	1	1	27	110	1	0	0	0	22	185	0	0	0	0	32

*Tabela 1(c.d.) – Wyniki badań dla próbek badawczych zapalników MD-8*Tabe

Lp.	Y	X1	X2	X3	X4	Lp.	Y	X1	X2	X3	X4	Lp.	Y	X1	X2	X3	X4
36	0	0	1	0	24	111	1	0	0	0	21	186	1	0	0	0	37
37	0	0	1	0	30	112	1	0	0	0	23	187	0	1	0	0	41
38	0	0	0	1	26	113	1	0	0	0	22	188	0	1	0	0	41
39	0	0	1	0	30	114	1	0	0	0	22	189	0	0	0	0	34
40	0	0	1	1	26	115	1	0	0	0	26	190	1	1	0	0	42
41	0	0	1	1	27	116	1	1	0	0	30	191	0	1	0	0	41
42	0	0	0	1	26	117	1	0	0	0	31	192	1	0	0	0	35
43	0	0	0	1	26	118	1	0	0	0	22	193	0	0	0	0	35
44	0	0	1	1	26	119	1	0	0	0	10	194	1	0	0	0	33
45	0	0	0	1	26	120	0	0	0	1	31	195	1	0	0	0	52
46	0	0	0	1	26	121	1	0	0	0	22	196	1	0	0	0	52
47	0	0	0	1	26	122	1	0	0	0	21	197	0	0	0	1	52
48	0	0	0	1	27	123	1	0	0	0	23	198	0	1	0	0	42
49	0	0	0	1	27	124	1	0	0	0	23	199	1	0	0	0	34
50	1	0	0	0	27	125	1	0	0	0	32	200	1	0	0	0	34
51	0	0	0	1	26	126	0	0	0	1	24	201	1	0	0	0	11
52	1	0	0	0	10	127	1	0	0	0	25	202	0	0	0	1	11
53	1	0	0	0	11	128	1	0	0	0	24	203	0	1	0	0	44
54	1	0	0	0	11	129	0	1	0	0	31	204	1	0	0	0	46
55	1	0	0	0	15	130	0	0	0	1	28	205	1	1	0	0	37
56	0	0	0	1	16	131	0	0	0	0	28	206	1	1	0	0	37
57	1	0	0	0	10	132	0	0	0	0	28	207	1	0	0	0	38
58	1	0	0	0	11	133	0	1	0	1	28	208	1	0	0	0	37
59	1	0	0	0	16	134	0	0	0	0	28	209	1	0	0	0	37
60	1	0	0	0	10	135	0	0	0	0	22	210	0	1	0	0	44
61	1	0	0	0	15	136	0	0	0	0	31	211	1	0	0	0	40
62	1	0	0	0	14	137	0	0	0	0	24	212	1	0	0	0	47
63	1	0	0	0	15	138	0	0	0	1	33	213	0	1	0	0	46
64	1	0	0	0	15	139	0	0	1	1	43	214	0	0	0	0	38
65	0	0	0	1	15	140	1	0	0	0	25	215	1	0	0	0	40
66	1	0	0	0	10	141	0	1	1	0	32	216	1	0	0	0	41
67	0	0	0	0	10	142	1	0	0	0	33	217	0	0	0	0	41
68	0	0	0	0	16	143	0	0	0	0	33	218	0	0	0	0	40
69	1	0	0	0	13	144	0	1	0	0	34	219	0	1	0	0	49
70	1	0	0	0	15	145	1	1	0	0	35	220	1	0	0	0	42
71	1	0	0	0	10	146	1	1	0	0	34	221	1	0	0	0	42
72	1	0	0	0	16	147	1	0	0	0	26	222	0	1	0	0	50
73	1	0	0	0	10	148	1	0	0	0	34	223	1	0	0	0	46
74	1	0	0	0	12	149	1	0	0	0	34	224	1	0	0	0	46
75	1	0	0	0	18	150	1	0	0	0	26						

Jako pierwszą zmienną niezależną do modelu wprowadzimy zmienną *X1*. Dzięki zastosowaniu oprogramowania [6] otrzymamy arkusz wyników widoczny na rysunku 1.

	Model: Regr. logistyczna (logit) N zer: 87 jedynek: 137 (MD-8) Zmn. zal.: Y Strata: Największe prawd. bł.średnkw.skal. Całkowita strata: 144,13462807 Chi2(1)=11,005 p=,00091	
N=224	Stała B0	X1
Ocena	0,6539265	-1,260062
Błąd standard.	0,1529211	0,3900999
t(222)	4,276235	-3,230102
p	0,00002824008	0,00142516
-95%CL	0,3525637	-2,028835
+95%CL	0,9552892	-0,4912896
Chi-kwadrat Walda	18,28618	10,43356
p	0,00001906313	0,001238711
Iloraz szans z.jedn.	1,923077	0,2836364
-95%CL	1,42271	0,1314886
+95%CL	2,599422	0,6118369
Iloraz szans zakr.		0,2836364
-95%CL		0,1314886
+95%CL		0,6118369

*Rys. 1 Arkusz wyników dla zmiennej X*1

Jak widać z przedstawionego arkusza wyników, zmienna *X1* oraz wartość stałej *B0* są wartościami istotnymi. Obliczony model może być dobrze dopasowany do analizowanych danych. Otrzymany model logistyczny przy rozpatrywaniu tylko tej zmiennej niezależnej *X1* miałby postać:

$$P\,(Y = 1) = \frac{e^{0,6539265-1,26006X1}}{1 + e^{0,6539265-1,26006X1}} \tag{8}$$

Przy tak przyjętych oznaczeniach (*Y = 1* oznacza przyjętą dodatnią decyzję), widoczny w arkuszu iloraz szans dla jednostkowej zmiany zmiennej *X1* równy 0,28 informuje nas, że prawdopodobieństwo uzyskania dodatniej decyzji podiagnostycznej przy występowaniu tego typu niezgodności, jest 0,28 razy większe niż przy jej niewystępowaniu.
Chcąc sprawdzić kolejne zmienne wzięte do analizy naszego modelu, budujemy dalej nasz hierarchiczny model logistyczny, dołączamy więc następną zmienną czyli wartości zmiennej niezależnej *X2*. Otrzymamy wówczas następujący arkusz wyników dla tych zmiennych widoczny na rysunku 2.
W przypadku dołączenia zmiennej *X2* widzimy, że wszystkie zmienne są także istotne statystycznie. Można stwierdzić, że dopasowanie naszego modelu do analizowanych danych empirycznych daje nam nowy model logistyczny istotny statystycznie. Otrzymany model logistyczny ma postać:

$$P\,(Y = 1) = \frac{e^{0,908676-1,470238X1-3,632685X2}}{1+e^{0,908676-1,470238X1-3,632685X2}} \tag{9}$$

N=224	Model: Regr. logistyczna (logit) N zer: 87 jedynek: 137 (MD-8) Zmn. zal.: Y Strata: Największe prawd. bł.średnkw.skal. Całkowita strata: 129,74622862 Chi2(2)=39,782 p=,00000		
	Stała B0	X1	X2
Ocena	0,9086756	-1,470238	-3,632685
Błąd standard.	0,1675084	0,3984131	1,04511
t(221)	5,424657	-3,690234	-3,475888
p	0,0000001519663	0,0002821941	0,000612812
-95%CL	0,5785574	-2,255413	-5,692342
+95%CL	1,238794	-0,6850625	-1,573029
Chi-kwadrat Walda	29,42691	13,61783	12,0818
p	0,00000005848621	0,000224419	0,0005098385
Iloraz szans z.jedn.	2,481035	0,2298709	0,02644508
-95%CL	1,783464	0,1048303	0,003371687
+95%CL	3,451448	0,5040587	0,2074161
Iloraz szans zakr.		0,2298709	0,02644508
-95%CL		0,1048303	0,003371687
+95%CL		0,5040587	0,2074161

Rys. 2 Arkusz wyników dla zmiennych $X1$ *i* $X2$

Kolejną zmienną dołączoną do modelu będzie zmienna niezależna $X3$. W przypadku tak obliczonego modelu dla trzech analizowanych zmiennych, model ten jest modelem również w pełni istotnym statystycznie.

N=224	Model: Regr. logistyczna (logit) N zer: 87 jedynek: 137 (MD-8) Zmn. zal.: Y Strata: Największe prawd. bł.średnkw.skal. Całkowita strata: 87,442386148 Chi2(3)=124,39 p=0,0000			
	Stała B0	X1	X2	X3
Ocena	1,769525	-2,178737	-3,740751	-5,155468
Błąd standard.	0,2360653	0,440248	1,090627	1,04277
t(220)	7,49591	-4,948886	-3,42991	-4,944011
p	0,000000000001594091	0,000001483405	0,0007208831	0,000001517187
-95%CL	1,304286	-3,04638	-5,890164	-7,210566
+95%CL	2,234764	-1,311094	-1,591338	-3,100371
Chi-kwadrat Walda	56,18867	24,49147	11,76428	24,44325
p	6,75407800E-14	0,0000007501813	0,0006045433	0,0000007691801
Iloraz szans z.jedn.	5,868063	0,1131844	0,02373627	0,005767778
-95%CL	3,685056	0,04753068	0,002766523	0,0007387388
+95%CL	9,344273	0,2695251	0,203653	0,0450325
Iloraz szans zakr.		0,1131844	0,02373627	0,005767778
-95%CL		0,04753068	0,002766523	0,0007387388
+95%CL		0,2695251	0,203653	0,0450325

Rys. 3 Arkusz wyników dla zmiennych $X1, X2$ *i* $X3$

Jak widać z arkusza wyników przedstawionego na rysunku 3 wszystkie zmienne są ważne tzn. stała $B0$ oraz zmienne $X1, X2\ i\ X3$ są wartościami istotnymi statystycznie. Można stwierdzić, że dołączenie analizowanej zmiennej $X3$ do naszego modelu nie wpływa na zmianę lub zepsucie dopasowania naszego modelu do analizowanych danych empirycznych. Otrzymany model logistyczny przy rozpatrywaniu trzech zmiennych będzie miał postać wyrażenia (10).

$$P\ (Y = 1) = \frac{e^{1,769-2,179X1-3,741X2-5,155X3}}{1+e^{1,769-2,179X1-3,741X2-5,155X3}} \qquad (10)$$

Zobaczmy więc, jak będzie wyglądał nasz model z ostatnią zmienną czyli zmienną $X4$. Jak wiemy pod tą zmienną kryje się wiek badanych partii zapalników typu MD-8. Stała $B0$ oraz zmienne $X1,\ X2,\ X3\ i\ X4$ są istotne statystycznie. Dołączenie tej zmiennej do modelu nie wpłynęło także ujemnie na dopasowanie modelu do analizowanych danych empirycznych. Na potwierdzenie tego faktu przedstawiam na rysunku 4 arkusz z wynikami dla wszystkich zmiennych jakie zostały analizowane w dopasowywanym modelu statystycznym.

	Model: Regr. logistyczna (logit) N zer: 87 jedynek: 137 (MD-8) Zmn. zal.: Y Strata: Największe prawd. bł.średnkw.skal. Całkowita strata: 84,156366921 Chi2(4)=130,96 p=0,0000				
N=224	Stała B0	X1	X2	X3	X4
Ocena	3,272861	-1,64573	-3,797206	-5,413678	-0,05531061
Błąd standard.	0,6770578	0,4787516	1,095729	1,058868	0,02186745
t(219)	4,833946	-3,437544	-3,465461	-5,112703	-2,529358
p	0,000002518482	0,0007024312	0,0006368156	0,0000006917673	0,01213089
-95%CL	1,938478	-2,58928	-5,95673	-7,500554	-0,09840821
+95%CL	4,607244	-0,7021797	-1,637683	-3,326802	-0,01221302
Chi-kwadrat Walda	23,36703	11,81671	12,00942	26,13974	6,397651
p	0,000001344733	0,0005877609	0,0005300165	0,0000003194078	0,01143191
Iloraz szans z.jedn.	26,38672	0,1928717	0,02243335	0,004455222	0,9461912
-95%CL	6,948166	0,07507406	0,002588362	0,000552778	0,9062789
+95%CL	100,2076	0,4955041	0,19443	0,03590773	0,9878613
Iloraz szans zakr.		0,1928717	0,02243335	0,004455222	0,09797471
-95%CL		0,07507406	0,002588362	0,000552778	0,01603238
+95%CL		0,4955041	0,19443	0,03590773	0,5987285

Otrzymany model logistyczny przy rozpatrywaniu czterech zmiennych będzie miał postać:

$$P\ (Y = 1) = \frac{e^{3,27-1,65X1-3,8X2-5,41X3-0,06X4}}{1+e^{3,27-1,65X1-3,8X2-5,41X3-0,06X4}} \qquad (11)$$

Na rysunkach 1÷4 w arkuszach wyników dla analizowanych zmiennych, oznaczonym jako Chi-kwadrat Walda mamy podane wartości testu czy też statystyki Walda sprawdzającej istotność parametrów, a w wierszu poniżej mamy wartości poziomu prawdopodobieństwa p z nim związane czyli poziom dobroci dopasowania.

W przypadku, kiedy podczas budowy powiększamy nasz model o jedną zmienną, a tak było w naszym przypadku, praktyczniejszy jest zawsze test Walda.

	Model: Regr. logistyczna (logit) N zer: 87 jedynek: 137 (MD-8) Zmn. zal.: Y Strata: Największe prawd. bł.średnkw.skal. Całkowita strata: 84,156366921 Chi2(4)=130,96 p=0,0000				
N=224	Stała B0	X1	X2	X3	X4
Ocena	3,272861	-1,64573	-3,797206	-5,413678	-0,05531061
Błąd standard.	0,6770578	0,4787516	1,095729	1,058868	0,02186745
t(219)	4,833946	-3,437544	-3,465461	-5,112703	-2,529358
p	0,000002518482	0,0007024312	0,0006368156	0,0000006917673	0,01213089
-95%CL	1,938478	-2,58928	-5,95673	-7,500554	-0,09840821
+95%CL	4,607244	-0,7021797	-1,637683	-3,326802	-0,01221302
Chi-kwadrat Walda	23,36703	11,81671	12,00942	26,13974	6,397651
p	0,000001344733	0,0005877609	0,0005300165	0,0000003194078	0,01143191
Iloraz szans z.jedn.	26,38672	0,1928717	0,02243335	0,004455222	0,9461912
-95%CL	6,948166	0,07507406	0,002588362	0,000552778	0,9062789
+95%CL	100,2076	0,4955041	0,19443	0,03590773	0,9878613
Iloraz szans zakr.		0,1928717	0,02243335	0,004455222	0,09797471
-95%CL		0,07507406	0,002588362	0,000552778	0,01603238
+95%CL		0,4955041	0,19443	0,03590773	0,5987285

Rys. 4 Arkusz wyników dla zmiennych $X1, X2, X3$ *i* $X4$

Otrzymujemy go obliczając kwadrat ilorazu szacowanego współczynnika regresji przez jego błąd standardowy. Jak widać z wyliczonych wartości tego testu, wartości te nieco się różnią. Natomiast poziomy prawdopodobieństwa są na poziomie wysoce istotnym i można zatem wyciągnąć wniosek, że wszystkie analizowane zmienne mają istotny wpływ na podejmowane decyzje podiagnostyczne.

Poziom prawdopodobieństwa p dla testu $t - Studenta$ towarzyszący testom istotności współczynników regresji jest dla wszystkich zmiennych na poziomie $p < 0{,}1$ czyli na poziomie niższym niż przyjęty poziom istotności. Jest on więc na poziomie wysoce istotnym i to oznacza, że parametry regresji są także istotne statystycznie.

Po określeniu naszego modelu do analizowanych zmiennych, obliczmy zatem wartość współczynnika R^2 dla tego przykładu.
Porównamy zatem wyliczone wartości mocy procesu predykcji dla czterech zaproponowanych modeli przedstawionych w artykule, z których każdy jest istotny statystycznie. Wyniki wartości parametru R^2 obliczonych za pomocą wzorów (2), (3), (5) i (6) oraz dzięki oprogramowaniu [6], które to oprogramowanie oblicza nam wartości logarytmów, przedstawione zostały w tabeli 2.

Tabela 2 - Wyniki obliczeń wartości R^2

Zmienne uwzględnione w analizowanym modelu	R^2 McFadden-a nie uwzględniający liczby zmiennych	R^2 McFadden-a z uwzględnioną liczbą zmiennych	R^2 Nagelkerke'a	R^2 Coxa-Snell-a
1	2	3	4	5
X1	0,036774	0,040115	0,065042	0,047944
X1 i X2	0,132929	0,139612	0,220754	0,162722
X1, X2 i X3	0,415638	0,425662	0,578072	0,426107
X1, X2, X3 i X4	0,437598	0,450964	0,600583	0,442700

Przeprowadzona krótka analiza obliczania wartości parametru R^2 świadczy więc o tym, że analizowane modele są dopasowane do danych empirycznych. Wyróżnikiem jednak w określeniu mocy dopasowania analizowanego modelu czy też mocy procesu predykcji jest obliczany parametr $\square^2$. Patrząc jednak na wartości tego parametru można stwierdzić, że najlepiej dopasowany jest model ze wszystkimi analizowanymi zmiennymi, ponieważ wartość parametru R^2 jest najwyższa. Jednakże nie należy do końca sugerować się tym wskaźnikiem, ponieważ jak wspomniałem wcześniej, niestety wadą tej sytuacji jest fakt, że w miarę zwiększania naszego modelu o kolejne zmienne niezależne, wzrastać też będzie wartość parametru R^2. Dlatego też większość statystyków za najbardziej wiarygodny parametr R^2 uważa współczynnik McFadden-a z uwzględnioną ilością analizowanych zmiennych, pomimo faktu, że współczynnik czy też parametr Nagelkerke uzyskał większe wartości. Bardziej zbliżone do współczynnika McFadden-a są natomiast obliczone wartości parametru R^2 Cox-Snell-a. W tabeli 2 wartości parametru *pseudo* R^2 McFadden-a z uwzględnioną liczbą zmiennych, zawarte są w kolumnie 3.

4. Wnioski

Przeprowadzona przykładowa analiza była próbą dopasowania modelu regresji logistycznej do danych empirycznych uzyskanych podczas badania zapalników dennych typu MD-8 w celu określenia wskaźnika *pseudo* R^2 czyli w celu określenia mocy procesu predykcji. Wyniki badań tych zapalników zostały otrzymane podczas ich badań laboratoryjnych. Analiza ta potwierdza opinię wielu statystyków, że współczynnik R^2 ma pewne wady. Są nimi głównie: wrażliwość na

liczbę zmiennych zastosowanych w modelu oraz brak możliwości osiągnięcia górnej granicy równej jeden.
W literaturze statystycznej można jednak spotkać określenie, że wartość parametru pseudo R^2 McFadden-a zawarta pomiędzy wartościami 0,2÷0,4, świadczy o dobrym dopasowaniu modelu. Na tej podstawie można stwierdzić, że dopasowanie naszego analizowanego modelu można uznać za satysfakcjonujące. W naszym przedstawionym przypadku wartość tego współczynnika dla czterech analizowanych zmiennych wyniosła $pseudo\ R^2 = 0{,}450964$, czyli jest wartością większą od zaleceń wielu statystyków. Moc procesu predykcji jest więc duża co oznacza, że analizowany model z czterema zmiennymi został dobrze dopasowany.
Przedstawiony w artykule sposób analizowania otrzymanych danych empirycznych pokazuje nam, że w wielu przypadkach podczas dopasowywania modelu, wszystkie wybrane zmienne mogą mieć istotny wpływ na dopasowanie szukanego modelu logistycznego do analizowanych danych empirycznych. Zmienne te oprócz tego, że są istotne mają również duży wpływ na określenie prawidłowej predykcji długości okresu bezpiecznej dalszej eksploatacji badanych elementów środków bojowych.
Reasumując, można stwierdzić, że pomimo faktu, iż obliczone wartości współczynnika R^2 Nagelkerke-a (tabela 2) mają wyższe wartości, w cale to nie oznacza, że moc procesu predykcji jest na tym poziomie dla analizowanego dopasowanego modelu. Ciągle jeszcze wybitni statystycy prowadzą rozmyślania oraz przeprowadzają szereg dowodów i analiz w temacie wyboru najbardziej wiarygodnego współczynnika R^2. Jak już wspomniałem wcześniej, wielu z nich po gruntownych badaniach oraz po szeregu przeprowadzonych dociekaniach naukowych, jest zdania, że wskaźnik $pseudo\ R^2$ McFadden-a jest tą wielkością najbardziej wiarygodną.
Również ważnym aspektem przy dopasowywaniu modelu do analizowanych danych empirycznych jest dobra znajomość specjalistycznego statystycznego oprogramowania komputerowego, którego użycie w obecnych czasach jest nieodzowne, aby wyznaczony cel osiągnąć i uzyskać prawidłową wartość szukanego parametru mocy procesu predykcji.

5. Literatura

[1] A. Stanisz – *Przystępny kurs statystyki* – Statsoft Polska, Kraków 2007 r.,

[2] *Wikipedia* – wolna encyklopedia,

[3] *Naukowiec.org.* – portal internetowy,

[4] P. Allison – *What's the Best R-Squared for Logistic Regression?* – Statistical Horizons, 2013 r.,

[5] M. Veall, K.. Zimmermann – *Pseudo-R2 Measures for Some Common Limited Dependent Variable Models* – 1996 r.,
[6] Statistica 12 – *Statsoft Polska 2012 r.* – oprogramowanie komputerowe,
[7] M. Rabiej – *Statystyka z programem Statistica* – Wydawnictwo Helion, Gliwice 2012 r.,
[8] S. Kot, J, Jakubowski, A. Sokołowski – *Statystyka* – Wydawnictwo Difin, Warszawa 2011 r.,
[9] G. S. Maddala – *Ekonometria* – PWN, Warszawa 2008 r.,
[10] Praca zbiorowa – Metodyka badań diagnostycznych amunicji – indeks N-5001b – 1985 r., Wojskowy Instytut Techniczny Uzbrojenia.

***Dr inż. Dariusz Ampuła**, absolwent Wojskowej Akademii Technicznej w zakresie Systemów Uzbrojenia Naziemnego. Po odbyciu rocznej praktyki zawodowej został pracownikiem naukowo-badawczym w Wojskowym Instytucie Technicznym Uzbrojenia w Zielonce. Stopień naukowy doktora uzyskał w 2006 roku w Instytucie Technicznym Wojsk Lotniczych. Autor i współautor prac z zakresu diagnostyki technicznej oraz niezawodności działania lądowych środków bojowych. W swoich analizach badawczych skupia się na problemach procesu predykcji badanych elementów amunicji lądowej.*