econometrics_model

Grzegorz Krochmal and Katarzyna Kry?ska

1. Introduction

In the business world it is important to know if the company you are dealing with is now and is going to be in the future in a good financial condition. Therefore, the way to predict the company's default has been the subject of interest for researchers for many years. The first important model to predict corporate bancruptcy was Altman's Z-Score model from 1968 (Altman, 1968). Since than many papers occured which were focusing this issue with different approaches. The techniques has evolved from Multidimensional Analysis proposed by Altman, through Logit (Ohlson, 1980) and Probit (Zmijewski, 1984) models, to modern solutions which incorporate highly developed Machine Learning techniques like Ensemble Boosted Trees(Tomczak et al., 2016). In our work we want to focus on Logit/Probit strategy of modelling company's default. Before the modelling part there is an obvious need to define main and the secondary hypothesis. In our case they are:

H0: *It is possible to predict company's default with specific financial indicators achieved by particular company*

H1: *Financial indicators can't be used as default predictors*

As it was stated before, it is very important for a market's health to be able to predict particular company's default with available financial indicators. It is crucial not only for investors who surely do not want to invest in companies which are going to go bankrupt but it is also really important for the whole country's economy. With the knowledge what financial indicators can predict company's default it would allow the policy makers to concentrate on this indicators and so on be able to prevent bankruptcies which would be especially important in case of major companies.

Some imporant works in the area of bankruptcy prediction were already mentioned. However this issue is deeply analysed in **Chapter 2**. In **Chapter 3** there is a description of a data set we have chosen for our modelling. Also in this part we introduce data processing and feature selection to properly conduct our study. In **Chapter 4** we make a choice between Probit and Logit model before running the model.

2. Literature Review

In the introduction few important works which challenged the problem of prediction of corporate bankruptcy were briefly mentioned. It all started with Altman's Multidemsional Analysis (Altman, 1968). Later many different ideas were introduce to solve the task of a proper prediction of bankruptcies. The works which are important for our analysis are ones which incorporate Logit (Ohlson, 1980) and Probit (Zmijewski, 1984). Obviously, it is important to state that those approaches are rather old and they do not keep up to the newest strategies. Currently, where the computing power is huge few modern techniques of bankruptcy prediction were proposed:

1. Rough Sets (Dimitras et al., 1999)
2. Evolutionary Programming (Zhang et al., 2013)
3. Ensemble Boosted Trees with Synthetic Features Generation (Tomczak et al., 2016)
4. Support Vector Machines (Shin et al., 2005)

However those methods can overcome the shortcomings of older approaches there are surely much more demanding. For example the SVM method which prediction power seems to be worse only than the Ensamble Boosted Trees method requires the function hand-tuning which makes it a tough tool for everyday business applications (Tomczak et al., 2016).
Coming back to the models which are the most important for this paper we are focusing on Ohlson's And Zmijewski's works. The Ohlson model is based on Logit. Beneath we present the list of variables Ohlson used in his analysis (Grice & Dugan, 2003):

- Y = The probability of membership in the bankrupt group based on a logistic function
- X1 = log (total assets/GNP price-level index)
- X2 = total liabilities/total assets
- X3 = working capital/total assets
- X4 = current liabilities/current assets
- X5 = one if total liabilities exceed total assets, zero otherwise
- X6 = net income/total assets
- X7 = funds provided by operations/total liabilities
- X8 = one if net income was negative for the last two years, zero otherwise
- X9 = measure of change in net income,1 and Y = overall index.

It is important to indicate that Ohlson selected mentioned variables in an arbitrary way, just choosing those which were most frequently mentioned in literature. It is some limitation which we will try to deal with in modelling part. This limitation is also the case with model proposed by Zmijewski. The researcher applied such variables:

- Y = overall index
- X1 = net income/total assets
- X2 = total debt/total assets
- X3 = current assets/current liabilities

They were all selected because of their importance in similar previous works. Additionaly, what has to be said about both models' shortcomings is the fact that the coefficients coming from the models are highly dependable on the data. So basically the models have to be retrained for every time period which is measured to obtain realiable results (Grice & Dugan, 2003). In our case it is not a significant problem but might be with huge datasets in business aplications. What might also be an important issue with those models is the fact that the dependent variables are selected arbitraty and their number is seriously limited. Nevertheless mentioned issues, both models still seem to be interesting and valuable. Now it is all the matter of the dataset how well those models are going to perform and eventually how we can improve them.

3. Data description and analysis

The analysed Data Set - Polish companies bankruptcy data - was created by Sebastian Tomczak and is about bankruptcy prediction of Polish companies. In our analysis we used financial rates from 5th year of the forecasting period that holds information about bankruptcy status after 1 year. We chose this dataset as it is the least imbalanced of all provided. This dataset contains 64 attributes and 5910 instances (financial statements) where 410 represent bankrupted firms and 5500 companies that did not bankrupt.

This dataset contains missing values. Firstly, we will look at columns to see which contain the most missing values.

	Number of missing values	Variable description
Attr37	2548	(current assets - inventories) / long-term liabilities
Attr27	391	profit on operating activities / financial expenses
Attr45	268	net profit / inventory
Attr60	268	sales / inventory
Attr24	135	gross profit (in 3 years) / total assets
Attr28	107	working capital / fixed assets

Attribut #37 contains 2548 values, which is almost 50% of number of rows, so we will dispose this variable. After that, we used Multivariate Imputation by Chained Equations to impute missing data where possible. MI imputes the missing values multiple times, resulting in

multiple full datasets. Then each dataset is analyzed and the results are combined.MI gives approximately unbiased estimates of all the estimates from the random error.

After these operations we have only 4.81% (284) of rows which contain missing data, so we will delete them.

4. Model estimation

In our analysis, we try to use models with Binary Dependent Variable, such as logit and probit. Our independent variables were chosen according to literature (Ohlson 1980). We tried to recreate all independent variables from Ohlson's research, leading us to eight independent variables:

- SIZE - Logarithm of total assets
- TLTA - Total liabilities divided by total assets
- WCTA - Working capital divided by total assets
- OENEG - One if total liabilities exceeds total assets, zero otherwise
- NITA - Net income divided by total assets
- INONE - One if net income was positive for the last year, zero otherwise
- CHSALES - Change in sales for the most recent period
- FUTL - Sum of gross profit and depreciation divided by total liabilities

Table 1: Summary of analysed variables

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
SIZE	5,626	4.201	0.793	0.006	3.699	4.691	9.698
TLTA	5,626	0.448	5.847	−430.870	0.266	0.662	72.416
WCTA	5,626	0.184	1.053	−72.067	0.043	0.410	0.994
OENEG	5,626	0.049	0.215	0	0	0	1
NITA	5,626	0.057	1.263	−32.052	0.004	0.115	87.459
INONE	5,626	0.788	0.409	0	1	1	1
FUTL	5,626	0.638	4.880	−221.330	0.075	0.623	217.330
CHSALES	5,626	2.538	102.132	−0.006	0.993	1.266	7,661.500
BANKRUPTCY	5,626	0.066	0.248	0	0	0	1

After wide literature review, we expect the sign of the coefficients to be as follows:

positive	negative	indeterminate
TLTA	WCTA	OENEG
INONE	NITA	
EBIT	CHSALES	
	SIZE	
	FUTL	

4.1. Estimation of linear probability model (OLS with White's robust matrix), logit model and probit model

The results of estimation are in Table 2. But before them there is one important issue which has to be discussed.

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

After running the code for probit and logit warnings occured. We have intentionally left them visible (first two are for probit, the last one for logit). The first one is just some computional warning of the function. It is about the number of iterations the glm probit is about to handle. The other two warnings (they are actually the same warning for different functions) indicate that we might be dealing with so called perfect or quasi perfect separation. To deal with this issue we could use Logistic Regression with Firth's Bias Reduction (Firth, 1993) which which would eliminate this. However there is a question whether we should try to deal with this at all. The perfect separation might not be the effect of some problem with our data sample but that just might be an expected outcome for a whole population. In our case we decided that perfect separation in our data can be a reflection of the whole population of companies. It is possible that the companies which go bankrupt have all the financial indicators worse than the healthy one competitors. That is why we decide to stay with standard logit/probit and omit the verion with Firth's Bias Reduction. However the Firth's approach is interesting and might be the case of interest in future studies in the area of corporate bankruptcies.

Table 2:

	Dependent variable:		
	BANKRUPTCY		
	OLS	*probit*	*logistic*
	(1)	(2)	(3)
SIZE	-0.031^{***}	-0.223^{***}	-0.427^{***}
	(0.004)	(0.038)	(0.076)
TLTA	0.003	0.083^{**}	0.147^{**}
	(0.004)	(0.033)	(0.068)
WCTA	-0.023^{***}	-0.540^{***}	-1.045^{***}
	(0.005)	(0.078)	(0.154)
OENEG	0.105^{***}	-0.028	-0.154
	(0.016)	(0.123)	(0.225)
NITA	0.012	0.348^{**}	0.612^{*}
	(0.017)	(0.152)	(0.320)
INONE	-0.140^{***}	-0.689^{***}	-1.391^{***}
	(0.009)	(0.072)	(0.147)
FUTL	-0.002^{***}	-0.192^{***}	-0.361^{***}
	(0.001)	(0.047)	(0.095)
CHSALES	-0.00001	-0.698^{***}	-1.450^{***}
	(0.00003)	(0.110)	(0.217)
Constant	0.306^{***}	0.642^{***}	1.568^{***}
	(0.018)	(0.185)	(0.359)
Observations	5,626	5,626	5,626
R^2	0.117		
Adjusted R^2	0.116		
Log Likelihood		$-1{,}060.814$	$-1{,}063.964$
Akaike Inf. Crit.		2,139.628	2,145.929
Residual Std. Error	0.233 (df = 5617)		
F Statistic	93.053^{***} (df = 8; 5617)		
Note:			*p<0.1; **p<0.05; ***p<0.01

4.2. Choice between logit and probit on the basis of information criteria

We know that our binary data is not balanced (zero values are much more frequent than one values), so we can use AIC criteria to choose between logit and probit model. The AIC criterion for probit is lower than for logistic model (look at Table 2), so in our analysis we will be using probit model.

4.3. Selection of significant variables; general-to-specific method to variables selection

We suspect that OENEG variable might not be signigicant. To test that, we will perform Waldtest for probit and logit model. The results of these test are in Table 3 and Table 4 respectively. P-value is above 0.05%, so we cannot reject the null hypothesis. However, removing OENEG from our model improves AIC cryteria. Therefore, we decided to remove it from the model

Table 3:

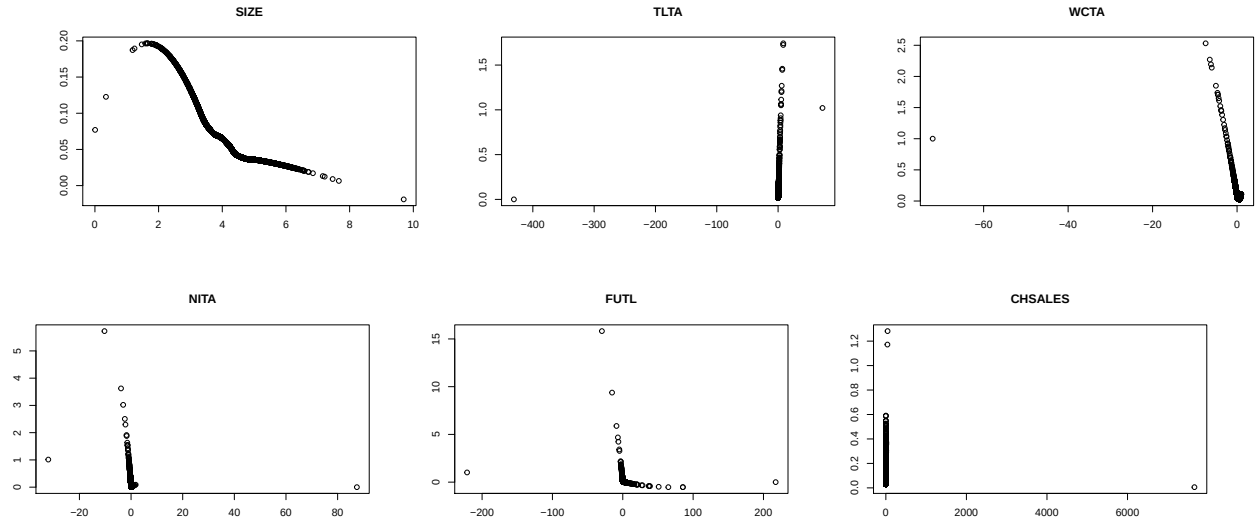
Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Res.Df	2	5,617.500	0.707	5,617	5,617.2	5,617.8	5,618
Df	1	−1.000		−1.000	−1.000	−1.000	−1.000
F	1	0.051		0.051	0.051	0.051	0.051
Pr(>F)	1	0.821		0.821	0.821	0.821	0.821

Table 4:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Res.Df	2	5,617.500	0.707	5,617	5,617.2	5,617.8	5,618
Df	1	−1.000		−1.000	−1.000	−1.000	−1.000
F	1	0.470		0.470	0.470	0.470	0.470
Pr(>F)	1	0.493		0.493	0.493	0.493	0.493

4.4. Transformation of variables

To check if any variables needs transformation, we will use loess plots for non-binary data.



Plot for SIZE shows that an assumption of a linear effect on the logit scale, is clearly not reasonable here. We will try to include a quadratic effect of SIZE in the model.

We also suspect that there might be interactions needed in our model. We excluded the list of possible pairs that have some scientific basis:

- CHSALES:INONE
- CHSALES:NITA
- NITA:INONE
- SIZE:TLTA

Beneath we present few likelihood ratio test where Model 1 is the one with some interaction included and Model 2 is our final probit model (it is described deeper in Chapter 4.5.)

```
## Likelihood ratio test
##
## Model 1: BANKRUPTCY ~ SIZE + TLTA + WCTA + OENEG + NITA + FUTL + SIZE2 +
##     CHSALES + CHSALES:INONE
## Model 2: BANKRUPTCY ~ SIZE + TLTA + WCTA + NITA + INONE + FUTL + SIZE2 +
##     CHSALES
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  10 -1063.3
## 2   9 -1060.3 -1 5.9341    0.01485 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Likelihood ratio test
##
## Model 1: BANKRUPTCY ~ SIZE + TLTA + WCTA + OENEG + INONE + FUTL + SIZE2 +
##     CHSALES + CHSALES:NITA
## Model 2: BANKRUPTCY ~ SIZE + TLTA + WCTA + NITA + INONE + FUTL + SIZE2 +
##     CHSALES
##   #Df  LogLik Df Chisq Pr(>Chisq)
## 1  10 -1064.4
## 2   9 -1060.3 -1 8.284      0.004 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Likelihood ratio test
##
## Model 1: BANKRUPTCY ~ SIZE + TLTA + WCTA + OENEG + NITA + FUTL + SIZE2 +
##     CHSALES + NITA:INONE
## Model 2: BANKRUPTCY ~ SIZE + TLTA + WCTA + NITA + INONE + FUTL + SIZE2 +
##     CHSALES
##   #Df   LogLik Df Chisq Pr(>Chisq)
## 1  10 -17913.7
## 2   9  -1060.3 -1 33707  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Likelihood ratio test
##
## Model 1: BANKRUPTCY ~ SIZE + TLTA + WCTA + OENEG + NITA + FUTL + SIZE2 +
##     CHSALES + NITA:INONE
## Model 2: BANKRUPTCY ~ SIZE + TLTA + WCTA + NITA + INONE + FUTL + SIZE2 +
##     CHSALES
##   #Df   LogLik Df Chisq Pr(>Chisq)
## 1  10 -17913.7
## 2   9  -1060.3 -1 33707  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

None of the interactions seem to be significant, so we will not include interactions in our model.

4.5. The final model

Table 5:

	Dependent variable:			
	BANKRUPTCY			
	OLS	*probit*	*logistic*	*probit*
	(1)	(2)	(3)	(4)
SIZE	−0.031***	−0.223***	−0.427***	0.017
	(0.004)	(0.038)	(0.076)	(0.244)
TLTA	0.003	0.083**	0.147**	0.079**
	(0.004)	(0.033)	(0.068)	(0.033)
WCTA	−0.023***	−0.540***	−1.045***	−0.540***
	(0.005)	(0.078)	(0.154)	(0.065)
OENEG	0.105***	−0.028	−0.154	
	(0.016)	(0.123)	(0.225)	
NITA	0.012	0.348**	0.612*	0.330**
	(0.017)	(0.152)	(0.320)	(0.149)
INONE	−0.140***	−0.689***	−1.391***	−0.683***
	(0.009)	(0.072)	(0.147)	(0.071)
FUTL	−0.002***	−0.192***	−0.361***	−0.195***
	(0.001)	(0.047)	(0.095)	(0.047)
SIZE2				−0.030
				(0.031)
CHSALES	−0.00001	−0.698***	−1.450***	−0.697***
	(0.00003)	(0.110)	(0.217)	(0.110)
Constant	0.306***	0.642***	1.568***	0.184
	(0.018)	(0.185)	(0.359)	(0.488)
Observations	5,626	5,626	5,626	5,626
R^2	0.117			
Adjusted R^2	0.116			
Log Likelihood		−1,060.814	−1,063.964	−1,060.301
Akaike Inf. Crit.		2,139.628	2,145.929	2,138.602
Residual Std. Error	0.233 (df = 5617)			
F Statistic	93.053*** (df = 8; 5617)			
Note:				*p<0.1; **p<0.05; ***p<0.01

Above we present the comparison of models tested in this analysis. Where the number four is our final model.

4.6. Marginal effects for the final model

We are going to calculate so called Marginal Effects at the Means

```
## Call:
## probitmfx(formula = BANKRUPTCY ~ SIZE + TLTA + WCTA + NITA +
##     INONE + FUTL + SIZE2 + CHSALES, data = ohl_data, atmean = TRUE)
##
## Marginal Effects:
##               dF/dx   Std. Err.       z   P>|z|
## SIZE     0.00014565  0.00206265  0.0706 0.94371
## TLTA     0.00066642  0.00037261  1.7885 0.07369 .
## WCTA    -0.00455981  0.00209548 -2.1760 0.02955 *
## NITA     0.00278515  0.00162432  1.7147 0.08641 .
## INONE   -0.01085253  0.00498089 -2.1788 0.02934 *
## FUTL    -0.00164798  0.00082124 -2.0067 0.04478 *
## SIZE2   -0.00025753  0.00028454 -0.9051 0.36543
## CHSALES -0.00588238  0.00179215 -3.2823 0.00103 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## dF/dx is for discrete change for the following variables:
##
## [1] "INONE"
```

We will consider only those marginal effects for variables which occured to be significant in regression (look Table 5). Due to the fact that the interpretation depends on the type of independent variable we will start from the only one discrete variable in our regression:

- INONE - 1 in this variable tells us that net income for previous year was positive. So the obtained MEM tells us that if net income was positive the probability of company's default falls by 8,4 percentage points. It is intuitive result.

Other variables are continous so the interpretation will be slightly different:

- TLTA - In this case we consider total liabilites divided by total assets. If TLTA increases by 0,1 then probability of default increases by 0,1 * 0,005 which is 0,05 percentage points. It is again intuitive conclusion that companies with higher liabilities are more likely to go bankrupt.

- WCTA - Working capital divided by total assets. In this case we find out that if WCTA increases by 0,1 then probability of default decreases by 0,1 * 0,036 which is 0,36 percentage points. Again this result is intuitive and goes with the predictions we made at the beginning of Chapter 4

- FUTL - Sum of gross profit and depreciation divided by total liabilities. If FUTL increases by 0,1 then probability of default decreases by 0,1 * 0,014 which is 0,14 percentage points. The higher gross profit the better for company's performance. This is also intuitive results

- CHSALES - Change in sales for the most recent period. If CHSALES increases by 0,1 then probability of default decreases by 0,1 * 0,0036 which is 0,036 percentage points. What we find out is that if the company achieved better performance in sales in this period than in the previous one it is less likely to go bankrupt. This is, again, an intuitive conclusion. What might be interesting and should be marked is the fact that this variable was significant in model itself but as a marginal effect got insignificant. However, due to the fact that the obtained conclusion fully complies with the theory and common sense it is still valuable for our analysis.

4.7. Calculation and interpretation of odds ratios

In this chapter we will analyse odds ratios of the model. We will also include 95% LR confidence intervals of odds.

```
## Waiting for profiling to be done...
```

```
##             odds_ratios     2.5 %    97.5 %
## (Intercept)   1.2018969 0.4705450 2.9483982
## SIZE          1.0173962 0.6498256 1.6308668
## TLTA          1.0821118 1.0319733 1.2510241
## WCTA          0.5827779 0.5210466 0.6908634
## NITA          1.3906943 1.1147792 1.7232015
## INONE         0.5050418 0.4416713 0.5715694
## FUTL          0.8227151 0.7579665 0.8894360
```

```
## SIZE2          0.9699652 0.9135867 1.0263973
## CHSALES        0.4982961 0.4069878 0.6082826
```

Similary as in 4.6 we fill focus only on variables which were significant in the Probit model. First we will analyse the only one dependent variable which is discrete:

- INONE - Odd ratio here tells us that the company with a positive net income for previouse year is 2 times less likely to go bankrupt than the one with a negative net income

Other variables are continous:

- TLTA - If TLTA increases by 1 the odds of company's default increases by 1,08 times
- WCTA - If WCTA increases by 1 the odds of company's default decreases by 1/0,58 = 1,72 times
- FUTL - If FUTL increases by 1 the odds of company's default decreases by 1/0,82 = 1,22 times
- CHSALES - If CHSALES incraeses by 1 the odds of company's default decreases by 1/0,5 = 2 times

4.8. Linktest

The next operation is to perform linktest. In this case we leave the code visible because it clarifies what the used variables are.

```
ohl_data$yhat = log(myprobit_final$fitted.values/(1-myprobit_final$fitted.values))
ohl_data$yhat2 = ohl_data$yhat^2
aux.reg = glm(BANKRUPTCY~yhat+yhat2, data=ohl_data, family=binomial(link="probit"))
summary(aux.reg)
```

```
##
## Call:
## glm(formula = BANKRUPTCY ~ yhat + yhat2, family = binomial(link = "probit"),
##     data = ohl_data)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -2.2147  -0.3305  -0.2247  -0.1472   3.5900
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.08695    0.08246  -1.054    0.292
## yhat         0.55354    0.06652   8.322   <2e-16 ***
## yhat2        0.01135    0.01357   0.836    0.403
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2729.1  on 5625  degrees of freedom
## Residual deviance: 2123.5  on 5623  degrees of freedom
## AIC: 2129.5
##
## Number of Fisher Scoring iterations: 11
```

It occurs that yhat2 is not significant. That indicates that the chosen model formula is correct.

4.9. Intepretation of R squared

When dealing with regression where the dependent variable is not continous the interpretation of R-squared is not straightforward as in case of linear regression. To get a wider understanding of how the model we prepared fits our data we will use three different pseudo R-squared measures: Count R^2 Nagelkerke/Cragg & Uhler R^2 and McFadden's

1. Count R^2 is equal

```
## [1] 0.9356559
```

We get a really high value of Count R^2 what tells us what percent of dependent variable's value we can predict with coefficients we got from the model. However the high value is optimistic, it is unfortunatelly the limitation of Count R^2 that if one value of dependent variable occurs much more often than another one the R^2 will surely be high (Strawiński). That is why some other measures should be introduced.

As the next value we decided to use Nagelkere's Pseudo R^2 because of its [0,1] interval which makes it comfortable to interpret:

```
## Nagelkerke
##  0.2667294
```

This value we can interpret that our model fully predicts 27% of the outcome.

The last one measure used is McFadden's:

```
##  McFadden
## 0.2229769
```

What does this value tell us about the predictive accuracy of our model? According to McFadden when his Pseudo R^2 is in range 0.2-0.4 it means a really good fit. And it is normal situation that this type of R^2 does not "behave" as well as this one from linear regression (McFadden, 1977)

Hypotheses verification

Tests

References

[1] xxx *The LaTeX Companion*. yyy, Reading, Massachusetts, 1993.