

# LOGISTIC REGRESSION IN IN-HOSPITAL MORTALITY MODELLING IN ACUTE MYOCARDIAL INFARCTION DATA



Václav Faltus<sup>1,2</sup>, Zdeněk Monhart<sup>3</sup>, Hana Grünfeldová<sup>1,4</sup>

<sup>1</sup> Centre of Biomedical Informatics, Prague, Czech Republic

<sup>2</sup> Dept. of Medical Informatics, Institute of Computer Science AS CR, v.v.i., Prague, Czech Republic

<sup>3</sup> Dept. of Medicine, Municipal Hospital Znojmo, Znojmo, Czech Republic

<sup>4</sup> Dept. of Medicine, Municipal Hospital Caslav, Caslav, Czech Republic

## Introduction

In this work we consider two approaches of modelling in-hospital mortality using logistic regression. The available data come from the pilot registry of myocardial infarction (MI) from the time period of 2003 - 2006. The registry comprises of all observations of acute myocardial infarction in registry hospitals in the Czech Republic. There were 7 participating hospitals from geographically different rural regions, some of them participating intermittently or not through the whole period. Totally there are 2659 observations of myocardial infarction - some likely to come from the identical individual but as there is no reason to treat the patients in a different way than when they got their previous myocardial infarction, the cases are assumed independent.

Our interest lies on in-hospital mortality depending on risk factors or their cumulation. The risk factors (RF) of interest are hypertension, hyperlipidaemia and diabetes mellitus. As other predictors we included age, smoking status, gender and additionally the year of observation. As people usually take some pharmacotherapy when they know they have the risk factor(s), the pharmacotherapy was not of interest in this work. Due to the missing data (with respect to the three risk factors of interests) we only used the 2435 complete observations. The year of observation can be regarded as a categorical or continuous variable and since there are certain discrepancies in the mortality rate during the whole period, we compare both methods. To compare the models we use the Akaike (AIC) and the Schwarz's Bayesian (BIC) information criteria.

	m1	m2	m3		n1	n2	n3
intercept	-1.619(0.209)	-6.844(0.635)	271.406(144.226)	intercept	-1.470(0.233)	-6.952(0.629)	308.713(140.243)
*	<.0001	<.0001	0.055	*	<.0001	<.0001	0.028
hyperlipidaemia	-0.707(0.150)	-0.495(0.155)	-0.484(0.154)	r3=1	-0.489(0.225)	-0.524(0.235)	-0.581(0.234)
*	<.0001	0.001	0.002	*	0.029	0.012	0.013
hypertension	-0.429(0.172)	-0.531(0.175)	-0.530(0.175)	r3=2	-0.660(0.223)	-0.708(0.230)	-0.697(0.229)
*	0.013	0.002	0.002	*	0.003	0.002	0.002
diabetes mellitus	0.585(0.152)	0.520(0.152)	0.527(0.152)	r3=3	-0.566(0.233)	-0.542(0.239)	-0.515(0.238)
*	<.0001	0.001	0.001	*	0.015	0.024	0.03
year2004	-0.453(0.182)	-0.414(0.186)	-	year2004	-0.480(0.180)	-0.426(0.185)	-
*	0.013	0.026	-	*	0.008	0.021	-
year2005	-0.684(0.217)	-0.636(0.222)	-	year2005	-0.711(0.215)	-0.672(0.220)	-
*	0.002	0.004	-	*	0.001	0.002	-
year2006	-0.365(0.207)	-0.355(0.211)	-	year2006	-0.417(0.205)	-0.410(0.210)	-
*	0.078	0.093	-	*	0.043	0.051	-
year	-	-	-0.139(0.070)	year	-	-	-0.150(0.070)
*	-	-	0.048	*	-	-	0.024
age	-	0.071(0.008)	0.072(0.008)	age	-	0.075(0.008)	0.076(0.008)
*	-	<.0001	<.0001	*	-	<.0001	<.0001
female gender	0.393(0.151)	-	-	female gender	0.422(0.149)	-	-
*	0.009	-	-	*	0.005	-	-
smoking	-0.673(0.221)	-	-	smoking	-0.774(0.219)	-	-
*	0.002	-	-	*	<.0001	-	-
Null Deviance	1491.2	1485.1	1485.1	Null Deviance	1491.2	1485.1	1485.1
Resid Deviance	1415.3	1335.3	1340.8	Resid Deviance	1445.6	1354	1359.3
Resid df	2403	2396	2396	Resid df	2403	2396	2396
AIC	1433.3	1351.3	1352.9	AIC	1463.6	1370	1371.3
BIC	1485.4	1397.6	1387.3	BIC	1515.7	1416.3	1406

Table 1: Considered logistic regression models with coefficients estimates with their standard errors, \*p-values, model characteristics (deviance and degrees of freedom) and information criteria (AIC, BIC).

Table 1 refers to six considered models used for modelling of in-hospital mortality in our data. In the first three models (m1, m2, m3) we used the three predictor variables themselves. The effects of hyperlipidaemia and hypertension on in-hospital mortality seem to be positive whereas the effect of diabetes mellitus is negative. The positive effect of hypertension and hyperlipidaemia can probably be explained by the fact that patients with these risk factors are usually monitored and the MI is successfully prevented or shifted to later age. In the second three models (n1, n2, n3) the estimated coefficients for the cumulative counts (variable r3) of risk factors are negative, meaning that the group with r3=0 has the highest in-hospital mortality. It is also the smallest RF group in our data, it contains only 13.0 % cases whereas there is 28.7 % in group r3=1, 34.3 % in group r3=2 and 24.0 % in group r3=3. In the group r3=0 there is 62.7 % of cases where the age of the patients is less than 70 years. In remaining RF groups this is 42.8 % (r3=1), 38.7 % (r3=2) and 36.7 % (r3=3). Higher proportion of younger people in the r3=0 group may also indicate genetic or other unknown factors evoking earlier myocardial infarction. The in-hospital mortality seems to decrease along with the increasing year of the MI observation (all models).

By looking at the AIC criterion the m2 model (n2 respectively) seems to be better than m3 (n3 respectively) but according to the BIC criterion we prefer the m3 model (n3 respectively) because it also reflects number of observations whereas the AIC does not. Predictors such gender and smoking status are only considered in the m1 model (n1 respectively). It is because the effect of smoking status and gender can be explained by the age variable as well. High proportion of men (Fig. 1) is smokers (vs. small proportion of smokers in women) and men are in average younger than women. The smoking status is not considered as the fourth risk factor since it is the least reliable data obtained only by asking the patient and believing his/her answer. Due to these facts only cumulative counts of maximum three risk factors are assumed. Since it seems not to be entirely true that the increasing number of risk factors increases the mortality, we used the variable as a factor and not as a continuous variable.

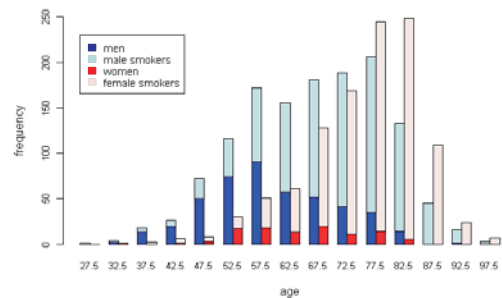


Figure 1: Histogram of age at the time of MI by gender and smoking status. Proportion of smokers is significantly higher in men than in women. Men are also significantly younger.

## Conclusion

In our work we present possible ways to model the in-hospital mortality, which is one of many important hospital quality measures. It was shown that using the cumulative count of risk factors instead of the separate risk factors themselves is of no negative effect on effectiveness modelling. However, the interpretation of results is different. Using risk factors as separate predictors helps to explain their effect on their own. Using the cumulative count helps to explain the cumulative effect of risk factors. The second mentioned approach is a little bit simpler in explaining to patients and thus probably more effective in clinical practice and counselling. A little bit surprising result of our work is the group with none or any known risk factors where the in-hospital mortality seems to exceed all remaining groups. This group also consists of much more MI cases, where the age of the patients is less than 70 years, when compared with other RF groups. This fact may indicate possible genetic or other unknown factors evoking earlier myocardial infarction.

## References

- Alan Agresti (2002) Categorical Data Analysis, 2nd Edition, Wiley.
- Monhart Z., Grünfeldová H., Ryšavá D., Veilmský T., Ballek J., Janský P., Faltus V. (2006) Myocardial infarction registry pilot project - results from the year 2004, Interv Akut Kardiol 5, 7377.
- R Development Core Team (2006) R: A language and environment for statistical computing, R Foundation for Statistical Computing.

## Contact

Mgr. Václav Faltus, MSc.  
EuroMISE Centre, Institute of Computer Science AS CR, v.v.i.  
Pod Vodárenskou věží 2, 182 07, Prague 8, Czech Republic  
faltus@euromise.cz, <http://www.euromise.cz>

## Abstract

Logistic regression is frequently used in analysing biomedical data, since they very often contain binary variables. Further, the situation where we have a large number of feasible predictors to include into the model arises very often. For binary response variables, the logit model is most suitable one but it faces very similar issues as there are for ordinary regression. The large number of possible predictors increases possible effects and interactions and makes data modelling more difficult. When modelling data, there are two competing goals: The model should be simple to interpret and use in practice, and it should be complex enough to fit the data well.

In this work we discuss different aspects of in-hospital mortality modelling in patients from the Myocardial Infarction Registry Pilot Project. The aim of this study was to compare diagnostic and therapeutic procedures in a population of all patients admitted with acute myocardial infarction to six municipal hospitals in the Czech Republic. The first aspect of modelling is to reduce the overfitted logit model through a proper choice of penalization criteria. The second aspect is to compare the model involving several binary variables - risk factors, with the model, which involves only the count of present risk factors despite the fact, that their individual effects on in-hospital mortality differ.