# Chapter 2
# Logistic Regression for Health Profiling

## 1 Summary

### *1.1 Background*

Logistic regression can be used for predicting the probability of an event in subjects at risk.

### *1.2 Methods and Results*

It uses log linear models of the kind of the one underneath (ln = natural logarithm, a = intercept, b = regression coefficient, x = predictor variable):

$$\text{"ln odds infarct} = a + b_1x_1 + b_2x_2 + b_3x_3 + \cdots\text{."}$$

A real data example was used of 1,000 subjects of different ages followed for 10 years for myocardial infarctions. Using the data, an exact risk could be calculated for individual future subjects.

### *1.3 Conclusions*

1. The methodology is currently an important way to determine, with limited health care sources, what individuals are at low risk and will, thus, be:

   (1) operated.
   (2) given expensive medications.
   (3) given the assignment to be treated or not.
   (4) given the "do not resuscitate sticker".
   (5) etc.

T.J. Cleophas and A.H. Zwinderman, *Machine Learning in Medicine*,
DOI 10.1007/978-94-007-5824-7_2, © Springer Science+Business Media Dordrecht 2013

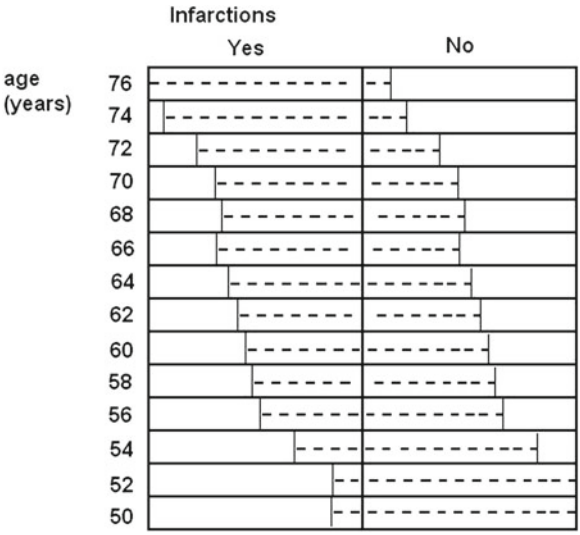


**Fig. 2.1** In a group of multiple ages the numbers of patients at risk of infarction is given by the *dotted line*

2. We must take into account that some of the predictor variables may be heavily correlated with one another, and the results may, therefore, be inflated.
3. Also, the calculated risks may be true for subgroups, but for individuals less so, because of the random error.

## 2 Introduction

Logistic regression can be used for predicting the probability of an event. For example, the odds of an infarction is given by the equation

$$\text{odds infarct in a group} = \frac{\text{number of patients with infarct}}{\text{number of patients without}}$$

The odds of an infarction in a group is correlated with age, the older the patient the larger the odds

According to Fig. 2.1 the odds of infarction is correlated with age, but we may ask how?

According to Fig. 2.2 the relationship is not linear, but after transformation of the odds values on the y-axis into log odds values the relationship is suddenly linear.

We will, therefore, transform the linear equation

$$y = a + bx$$


**Fig. 2.2** Relationships between the odds of infarction and age

into a log linear equation ( ln = natural logarithm)

$$\ln \text{odds} = a + b\ x\ \left(x = \text{age}\right).$$

## 3 Real Data Example

Our group consists of 1,000 subjects of different ages that have been observed for 10 years for myocardial infarctions. Using SPSS statistical software, we command binary logistic regression

$$\begin{aligned}&\text{dependent variable infarction yes/no } \left(0/1\right)\\ &\text{independent variable age}\end{aligned}$$

The program produces a regression equation:

$$\ln \text{odds} = \ln\ \frac{\text{pts with infarctions}}{\text{pts without}} = a + bx$$

$a = -9.2$
$b = 0.1$ ($SE = 0.04$; $p < 0.05$)

The age is, thus, a significant determinant of odds infarction (which can be used as surrogate for risk of infarction).

Then, we can use the equation to predict the odds of infarction from a patient's age:

$$\text{Ln odds 55 years} = -9.2 + 0.1 \,.\, 55 = -4.82265$$
$$\text{odds} = 0.008 = 8 / 1{,}000$$

$$\text{Ln odds 75 years} = -9.2 + 0.1 \,.\, 75 = -1.3635$$
$$\text{odds} = 0.256 = 256 / 1{,}000$$

Odds of infarction can, of course, more reliably be predicted from multiple x-variables. As an example, 10,000 pts are followed for 10 years, while infarctions and baseline-characteristics are registered during that period.

| | |
|---|---|
| dependent variable | infarction yes/no |
| independent variables (predictors) | gender |
| | age |
| | Bmi (body mass index) |
| | systolic blood pressure |
| | cholesterol |
| | heart rate |
| | diabetes |
| | antihypertensives |
| | previous heart infarct |
| | smoker |

The data are entered in SPSS, and it produces b-values (predictors of infarctions)

| | b-values | p-value |
|---|---|---|
| 1. Gender | 0.6583 | <0.05 |
| 2. Age | 0.1044 | " |
| 3. Bmi | –0.0405 | " |
| 4. Systolic blood pressure | 0.0070 | " |
| 5. Cholesterol | 0.0008 | " |
| 6. Heart rate | 0.0053 | " |
| 7. Diabetes | 1.2509 | <0.10 |
| 8. Antihypertensives | 0.3175 | <0.050 |
| 9. Previous heart infarct | 0.8659 | <0.10 |
| 10.Smoker | 0.0234 | <0.05 |
| a-value | –9.1935 | " |

It is decided to exclude predictors that have a p-value > 0.10.
The underneath regression equation is used

$$\text{"ln odds infarct} = a + b_1x_1 + b_2x_2 + b_3x_3 + \ldots\text{."}$$

to calculate the best predictable y-value from every single combination of x-values.

For instance, for a subject with the following characteristics (= predictor variables)

- Male ($x_1$)
- 55 years of age ($x_2$)
- cholesterol 6.4 mmol/l ($x_3$)
- systolic blood pressure 165 mmHg ($x_4$)
- antihypertensives ($x_5$)
- dm ($x_6$)
- 15 cigarettes/day ($x_7$)
- heart rate 85 beats/min ($x_8$)
- Bmi 28.7 ($x_9$)
- smoker ($x_{10}$)

the calculated odds of having an infarction in the next 10 years is the following:

| | b-values | x-values | |
|---|---|---|---|
| Gender | 0.6583 . | 1 (0 or 1) | = 0.6583 |
| Age | 0.1044 . | 55 | = 5.742 |
| BMI | –0.0405 . | 28.7 | = .. |
| Blood pressure | 0.0070 . | 165 | = |
| Cholesterol | 0.0008 . | 6.4 | = |
| Heart rate | 0.0053 . | 85 | = |
| Diabetes | 1.2509 . | 1 | = |
| Antihypertensives | 0.3175 . | 1 | = |
| Previous heart inf | 0.8659 . | 0 | = |
| Smoker | 0.0234 . | 15 | = |
| a-value | | | = –9.1935 + |
| | | Ln odds infarct | = –0.5522 |
| | | odds infarct | = 0.58 = 58/100 |

The odds is often interpreted as risk. However, the true risk is a bit smaller than the odds, and can be found by the equation

$$\text{risk event} = 1/\left(1+1/\text{odds}\right)$$

If odds of infarction = 0.58, then the true risk of infarction = 0.37.

## 4 Discussion

The above methodology is currently an important way to determine, with limited health care sources, what individuals will be:

1. operated.
2. given expensive medications.

**Table 2.1** Examples of predictive models where multiple logistic regression has been applied

| Dependent variable (odds of event) | Independent variables (predictors) |
|---|---|
| 1. TIMI risk score [1] | |
| Odds of infarction | Age, comorbidity, comedication, riskfactors |
| 2. Car producer (Strategic management research) [2] | |
| Odds of successful car | Cost, size, horse power, ancillary properties |
| 3. Item response modeling (Rasch models for computer adapted tests) [3] | |
| Odds of correct answer to three questions of different difficulty | Correct answer to three previous questions |

3. given the assignment to be treated or not.
4. given the "do not resuscitate sticker".
5. etc.

We need a large data base to obtain accurate b-values. This logistic model for turning the information from predicting variables into probability of events in individual subjects is being widely used in medicine, and was, for example, the basis for the TIMI (Thrombolysis In Myocardial Infarction) prognostication risk score. However, not only in medicine, also in strategic management research, psychological tests like computer adapted tests, and many more fields it is increasingly observed (Table 2.1). With linear regression it is common to provide a measure of how well the model fits the data, and the squared correlation coefficient $r^2$ is mostly applied for that purpose. Unfortunately, no direct equivalent to $r^2$ exists for logistic, otherwise called loglinear, models. However, pseudo-R2 or R2-like measures for estimating the strength of association between predictor and event have been developed.

Logistic regression is increasingly used for predicting the risk of events like cardiovascular events, diagnosis of cancer, deaths etc. Limitations of this approach have to be taken into account. It uses observational data and may, consequently, give rise to serious misinterpretations of the data:

1. The assumption that baseline characteristics are independent of treatment efficacies may be wrong.
2. Sensitivity of testing is jeopardized if the models do not fit the data well enough.
3. Relevant clinical phenomena like unexpected toxicity effects and complete remissions can go unobserved.
4. The inclusion of multiple variables in regression models raises the risk of clinically unrealistic results.

As another example, a cohort of postmenopausal women is assessed for the risk of endometrial cancer. The main question is: what are the determinants of endometrial cancer in a category of females. The following logistic model is used:

y-variable = ln odds endometrial cancer
$x_1$ = estrogene consumption short term

$x_2$ = estrogene consumption long term
$x_3$ = low fertility index
$x_4$ = obesity
$x_5$ = hypertension
$x_6$ = early menopause

lnodds endometrial cancer = $a + b_1$ estrogene data + $b_2$.... + $b_6$ early menopause data
The odds ratios for different x-variables are defined, e.g., for:

$x_1$ = chance cancer in consumers of estrogene/non-consumers
$x_3$ = chance cancer in patients with low fertility/their counterparts
$x_4$ = chance cancer in obese patients/their counterparts etc.

| risk factors | regression coefficient(b) | standard error | p-value | odds ratio ($e^b$) |
|---|---|---|---|---|
| 1. estrogenes short | 1.37 | 0.24 | <0.0001 | 3.9 |
| 2. estrogenes long | 2.60 | 0.25 | <0.0001 | 13.5 |
| 3. low fertility | 0.81 | 0.21 | 0.0001 | 2.2 |
| 4. obesity | 0.50 | 0.25 | 0.04 | 1.6 |
| 5. hypertension | 0.42 | 0.21 | 0.05 | 1.5 |
| 6. early menopause | 0.53 | 0.53 | ns | 1.7 |

The data were entered in the software program, which provided us with the best fit b-values. The model not only showed a greatly increased risk of cancer in several categories, but also allowed to consider that the chance of cancer if patients consume estrogens, suffer from low fertility, obesity, and hypertension might have an increased risk as large as $= e^{b2+b3+b4+b5} = 75.9 = 76$ fold. This huge chance is, of course, clinically unrealistic! We must take into account that some of these variables must be heavily correlated with one another, and the results are, therefore, largely inflated. In conclusion, logistic regression is an adequate tool for exploratory research, the conclusions of which must be interpreted with caution, although they often provide scientifically highly interesting questions. This should be kept in mind when using it for health profiling in individuals. Also the calculated risk may be true for subgroups, but for individuals less so, because of the random error.

## 5 Conclusions

Logistic regression can be used for predicting the probability of an event. It uses log linear models of the kind of the one underneath:

$$„"\ln \text{odds infarct} = a + b_1x_1 + b_2x_2 + b_3x_3 + \cdots."$$

The methodology is currently an important way to determine, with limited health care sources, what individuals will be:

1. operated.
2. given expensive medications.
3. given the assignment to be treated or not.
4. given the "do not resuscitate sticker".
5. etc.

We must take into account that some of these variables must be heavily correlated with one another, and the results are, therefore, largely inflated. Also the calculated risks may be true for subgroups, but for individuals less so, because of the random error.

## References

1. Antman EM, Cohen M, Bernink P, McGabe CH, Horacek T, Papuches G, Mautner B, Corbalan R, Radley D, Braunwald E (2000) The TIMI risk score for unstable angina pectors, a method for prognostication and therapeutic decision making. J Am Med Assoc 284:835–842
2. Hoetner G (2007) The use of logit and probit models in strategic management research. Strateg Manag J 28:331–343
3. Rudner LM Computer adaptive testing. http://edres.org/scripts/cat/catdemo.htm. Accessed 18 Dec 2012

http://www.springer.com/978-94-007-5823-0

Machine Learning in Medicine

Cleophas, T.J.; Zwinderman, A.H.
2013, XV, 265 p. 44 illus., Hardcover
ISBN: 978-94-007-5823-0