How to Deal with Continuous and Dichotomic Outcomes in Epidemiological Research: Linear and Logistic Regression Analyses

Giovanni Tripepi^a Kitty J. Jager^b Vianda S. Stel^b Friedo W. Dekker^{b, c}
Carmine Zoccali^a

^aCNR-IBIM, Clinical Epidemiology and Physiopathology of Renal Diseases and Hypertension, Reggio Calabria, Italy;

^bERA-EDTA Registry, Department of Medical Informatics, Academic Medical Center, University of Amsterdam, Amsterdam, and ^cDepartment of Clinical Epidemiology, Leiden University Medical Centre, Leiden, The Netherlands

Key Words

Confounding • Linear regression analyses • Logistic regression analysis • Multiple regression analysis

Abstract

Because of some limitations of stratification methods, epidemiologists frequently use multiple linear and logistic regression analyses to address specific epidemiological questions. If the dependent variable is a continuous one (for example, systolic pressure and serum creatinine), the researcher will use linear regression analysis. Otherwise, if the dependent variable is dichotomic (for example, presence/absence of microalbuminuria), one could use logistic regression analysis. In both linear and logistic regression analyses the independent variables may be either continuous or categorical. In this paper we will describe linear and logistic regression analyses by discussing methodological features of these techniques and by providing clinical examples and guidance (syntax) for performing these analyses by commercially available statistical packages. Furthermore, we will also focus on the use of multiple linear and logistic regression analyses to control for confounding in etiological research.

Copyright © 2011 S. Karger AG, Basel

Introduction

In two previous papers [1, 2] we described two stratification methods to control for confounding in epidemiological research. Because of some limitations of stratification methods [1, 2], epidemiologists frequently use multiple linear and logistic regression analyses to address specific epidemiological questions. In fact, in epidemiological research we could be interested in investigating the relationship between one or more risk factors and a continuous outcome variable (like systolic arterial pressure, serum creatinine or microalbuminuria) or a dichotomous outcome variable (presence/absence of hypertension, of microalbuminuria, etc.). If the dependent variable is a continuous one, the researcher will use linear regression analysis. Otherwise, if the dependent variable is dichotomic, one could use logistic regression analysis. In both linear and logistic regression analyses the independent variables may be either continuous or categorical. In general, regression analyses (either linear or logistic) are used to describe the dependence of the outcome variable (also called dependent variable) from one (or more) explanatory variables (or independent variables). In this paper we will describe linear and logistic regres-

Fig. 1. Relationship between serum glucose and 24-hour microalbuminuria in 98 hypertensive patients with 24-hour microalbuminuria ranging from 60 to 200 mg/24 h. The concept of 'residual' is described graphically as the distance (vertical dotted line) between each observed value and the regression line (see text for more details).

sion analyses by discussing methodological features of these techniques and by providing clinical examples and guidance (syntax) for performing these analyses by commercially available statistical packages. Furthermore, we will also focus on the use of multiple linear and logistic regression analyses to control for confounding [3] in etiological research.

Linear Regression Analysis

Example 1

Here we consider a study including 98 hypertensive patients with 24-hour microalbuminuria ranging from 60 to 200 mg/24 h. The aim of the study was to assess the relationship between fasting serum glucose and microalbuminuria by linear regression analysis. In linear regression analysis the explanatory variable is plotted on the X scale and the outcome variable on the Y scale. Each dot in the graph represents an individual and it is identified by a pair of values: the value of serum glucose (X scale) and the corresponding value of microalbuminuria (Y scale). Figure 1 shows that 24-hour microalbuminuria increases in parallel with serum glucose indicating a linear relationship between the two variables. The linear dependence of 24-hour microalbuminuria from serum glucose levels was assessed by calculating the increase in 24-hour microalbuminuria triggered by 1 mg/dl increase in se-

rum glucose. This information was obtained by using the equation derived from the regression line describing the 24-hour microalbuminuria-glucose relationship. In general terms, a regression line can be calculated by the equation:

$$E(y) = \beta_0 + \beta_1 x$$

where $E(y)$ is the estimated or predicted value of the dependent variable Y (24-hour microalbuminuria); β_0 is the intercept; β_1 is the regression coefficient, and x is a given value of the explanatory or independent variable (serum glucose).

The intercept (β_0) is the theoretical value of Y when X equals 0 (fig. 1). The regression coefficient (β_1) is the estimated increase in the dependent variable (Y) per 1 unit increase in the independent variable (X), in other words the slope of the regression line (fig. 1). The method used to estimate the intercept and the regression coefficient is the least-squares method [4]. The differences between the observed data points and the predicted values in the regression line are called residuals (see vertical dotted line in fig. 1). The least-squares method consists of finding the parameters (β_0 and β_1) that minimize the sum of the squares of these residuals.

The equation describing the regression line of the microalbuminuria-glucose link as provided by the statistical software is:

$$\text{estimated 24-hour microalbuminuria} = 49 + 0.75 \cdot \text{glucose (mg/dl)}$$

A regression coefficient of 0.75 means that for each 1 g/dl increase in serum glucose there is a parallel increase of 0.75 mg/24 h in microalbuminuria [e.g. for 20 mg/dl increase in serum glucose there is an average increase of 15 mg/24 h (i.e. 0.75×20) in microalbuminuria]. In this perspective, it is important to realize that the interpretation of the values of the regression coefficients is strictly dependent on the units of measurement of both the dependent and the independent variable. A positive regression coefficient indicates a positive relationship between risk factor and outcome variable (direct relationship) and a negative regression coefficient indicates a negative one (inverse relationship). The value of the intercept (49 mg/24 h) corresponds to the theoretical value of 24-hour microalbuminuria when serum glucose is 0 (fig. 1). The computation of the intercept is useful for predictive purposes because it can be used, together with the regression coefficient, to predict the estimated value of 24-hour microalbuminuria for a given individual, of whom we know the corresponding serum glucose concentration. For ex-

ample, the estimated value of 24-hour microalbuminuria for an individual having a serum glucose of 150 mg/dl (see dot indicated by the arrow in fig. 1) can be calculated by resolving the equation:

$$\text{estimated 24-hour microalbuminuria} = 49 + 0.75 \cdot 150 = 161 \text{ mg/24 h}$$

Thus, by using the regression line constructed in our sample, we could normally predict a 24-hour microalbuminuria of 161 mg/24 h for an individual having a serum glucose of 150 mg/dl. For this individual, the residual is calculated as the difference between the observed (152 mg/24 h) and the estimated value of 24-hour microalbuminuria (161 mg/24 h), which would be 9 mg/24 h. By repeating this calculation for all observed and predicted values we obtain a distribution of residuals, these latter ranging from -54 to +76 mg/24 h. Such a wide distribution indicates that in this particular case serum glucose does not accurately predict 24-hour microalbuminuria at an individual level. As a consequence of the wide range of residuals, the correlation coefficient (r), which defines how much the linear model we apply to describe the serum glucose to 24-hour microalbuminuria relationship is compatible with a straight line, is rather low ($r = 0.33$, see fig. 1). Beyond its predictive purposes, the analysis of residuals is particularly relevant for testing the three assumptions underlying the linear regression analysis that rest on three statements: (1) the relationship between the two variables is linear; (2) at each value of the independent variable (X scale) there is a correspondent set of normally distributed values of the dependent variable (Y scale), and (3) the standard deviation of this set of values is the same for each value of the independent variable [5]. If all these assumptions are true, the residuals should be normally distributed. In our instance, the residuals of the 24-hour microalbuminuria-glucose link have an approximately normal distribution (data not shown), indicating that the data distribution in the sample meets all above-mentioned criteria. When there is a violation of the normality assumption related to residual analysis, a means to solve the problem is the mathematical transformation of the dependent/independent variables. The transformation to be used depends on the degree of the deviation from normality. If we deal with a negatively skewed distribution of the dependent/independent variable [that is, the values tend to cluster toward the higher end of the distribution plot (that is, the higher numbers)], a square root transformation of the variable is often the best. Otherwise, if the variable is positively skewed distributed [that is, the values tend to cluster toward the lower end of

the distribution plot (that is, the lower numbers)], a log transformation usually solves the problem. If all these methods do not solve the problem, the dichotomization of the variable(s) can be a valid alternative. A detailed description of the methods used to account for the violation of the normality assumption is given elsewhere [5, pp. 251–252].

Linear and Logistic Regression Analysis as Tools to Control for Confounding

In a previous paper of this series [3] we discussed that ‘confounding’ may distort the true effect of a given exposure on a specific outcome. Multiple linear regression analysis allows estimation of the linear effect of an explanatory variable on a given outcome variable (Y) after controlling for the confounding effect of other variables (for example $x_2, x_3 \dots x_n$). The corresponding equation of the multiple linear regression model is:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots \beta_n x_n$$

where $E(y)$ is the estimated or predicted value of Y ; β_0 is the intercept (i.e. the value of Y when x_1, x_2 and x_3 are zero); $\beta_1, \beta_2, \beta_3$ and β_n are the regression coefficients of x_1, x_2, x_3 and x_n .

In the previous example we described the relationship between 24-hour microalbuminuria and serum glucose in hypertensive patients and found that the two variables were directly related. Now, we analyze the effect of serum glucose on 24-hour microalbuminuria by considering the confounding effect of age, a variable that resulted to be directly related to both 24-hour microalbuminuria ($r = 0.29, p = 0.001$) and serum glucose ($r = 0.23, p = 0.02$) (fig. 2). We consider age as a potential confounder because it meets criteria set for the definition of confounder [3]. In fact, age influences both 24-hour microalbuminuria (the outcome variable) and serum glucose (the explanatory variable); it cannot be considered as an effect of exposure and we assume that age is not in the causal pathway between the exposure (serum glucose) and the outcome (24-hour microalbuminuria). After introducing age into the multiple linear model, the regression line provided by the computer output is:

$$\text{estimated 24-hour microalbuminuria} = 37 + 0.63 \cdot \text{glucose (mg/dl)} + 0.56 \cdot \text{age (years)}$$

A 0.63 regression coefficient for serum glucose means that for each 1 mg/dl increase in this variable there is a 0.63 mg/24 h increase in microalbuminuria and this estimate is adjusted for the confounding effect of age. Comparing the adjusted effect (0.63) versus the unadjusted ef-

Fig. 2. Relationship between age and 24-hour microalbuminuria or serum glucose in 98 hypertensive patients with 24-hour microalbuminuria ranging from 60 to 200 mg/24 h. Data are Pearson correlation coefficient and p value.

Table 1. Association between smoking and development of macroalbuminuria in 100 diabetic patients

	Proportion of smokers (p)	Odds of smoking [p/(1 - p)]	Logit (ln odds)
Patients with macroalbuminuria (n = 53)	0.509	1.037	0.036
Patients without macroalbuminuria (n = 47)	0.277	0.383	-0.959

fect reported above (0.75) we see that age was indeed a confounder, here as adjustment for age changed the effect of glucose on 24-hour microalbuminuria.

Number of Covariates into the Multiple Linear Regression Analysis

A critical question is how many covariates can be entered into a multiple linear regression analysis. The number of covariates allowed depends on the sample size. A practical rule is to include 1 covariate for every 10 observations [5, pp. 389–390]. Thus, if we are to construct a model based on 10 variables, the general rule makes necessary a sample size of 100 individuals.

Logistic Regression Analysis

Linear regression analysis demands that the dependent variable be continuous. However, many clinical or epidemiological variables are dichotomic in nature: for example, a patient may or may not be affected by a given

disease or he can die or survive during a given time period. Logistic regression analysis is a statistical technique that describes the relationship between an independent variable (either continuous or not) and a dichotomic-dependent variable (or dummy variable; i.e. a variable with only two possible values: 0 = outcome absent and 1 = outcome present). Logit transformation (see below) is the fundamental mathematical step underlying this analysis.

Example 2

We consider a hypothetical study investigating the relationship between cigarette smoking (0 = nonsmoker; 1 = smoker) and the transition from microalbuminuria to macroalbuminuria (dependent variable) in a series of 100 diabetic patients treated by hypoglycemic agents and with an average microalbuminuria of 200 ± 40 mg/24 h at baseline. Another aim of the study is to investigate whether the link between smoking and occurrence of macroalbuminuria is affected by the confounding effect of age, a variable that we assume is not involved in the causal pathway between the exposure (smoking) and the outcome (macroalbuminuria). All patients were followed up for 5 years and during the follow-up 53 patients out of 100 (53%) developed macroalbuminuria. Here we are not interested in the time to event analysis but only in the occurrence of macroalbuminuria. The proportion of smokers in patients with and without macroalbuminuria is given in table 1.

The concept of odds is described in a previous article of this series [6].

The proportion of smokers was about 2 times higher in patients with macroalbuminuria (0.509, i.e. 50.9%) than in those without (0.277, i.e. 27.7%) (table 1).

Odds, Odds Ratio and Logit

The odds of smoking (third column) were calculated by the standard formula:

$$\text{odds} = [p/(1 - p)]$$

where p is the proportion of smokers in patients with and without macroalbuminuria (table 1).

According to the formula, in the group of patients with macroalbuminuria the odds of smoking were:

$$\text{odds} = 0.509/(1 - 0.509) = 1.037$$

In the group of patients without macroalbuminuria the odds of smoking were:

$$\text{odds} = 0.277/(1 - 0.277) = 0.383$$

Thus, the odds ratio of smoking between patients with and without macroalbuminuria will be the ratio between the two odds:

$$\text{odds ratio} = 1.037/0.383 = 2.71$$

The next step was the calculation of logit (or logistic) transformation of the odds of smoking (table 1; last column). The logit is the natural logarithm ($\ln$) of the odds:

$$\text{logit} = \ln[p/(1 - p)]$$

For example, the logit transformation of the odds of smoking in patients with macroalbuminuria was:

$$\text{logit} = \ln(1.037) = 0.036$$

As for the linear regression analysis, also in logistic regression analysis the relationship between outcome occurrence and explanatory variable is described by an equation:

$$\text{logit } y = \beta_0 + \beta_1 x$$

In the above equation the intercept (β_0) is the value of the natural logarithm of the odds of a given outcome when exposure equals 0 and the regression coefficient (β_1) is the logarithm of the odds of a given outcome when exposure is present. In the logistic regression analysis, the regression coefficients are calculated by using the maximum likelihood method [5, pp. 639–655]. The regression coefficients are directly provided by the output of the statistical software.

In our case the equation of the logistic model is:

$$\text{logit of macroalbuminuria } (y) = -0.27 + 0.999 \cdot \text{smoking} \\ (0 = \text{no}; 1 = \text{yes})$$

To estimate the increase in the odds of macroalbuminuria in smokers as compared to that in nonsmokers we

calculated the inverse operation of logit transformation, i.e. the antilogarithm of the regression coefficient, thus obtaining the odds ratio. In other words, we computed the odds ratio by exponentiating the base of the natural logarithm ($e = 2.7183$) to the regression coefficient (β_1): $2.7183\beta_1$. Therefore, the odds ratio corresponding to a regression coefficient of 0.999 is:

$$\text{odds ratio} = e\beta_1 = 2.7183^{0.999} = 2.71$$

An odds ratio of 2.71 means that the odds of smoking are 2.71 times higher in patients with macroalbuminuria than in those without macroalbuminuria, suggesting an association of smoking with the transition from micro- to macroalbuminuria in diabetic patients. An important property of the odds ratio is that the odds ratio of exposure (smoking) equals the odds ratio of the outcome (transition to macroalbuminuria). Therefore, we can also conclude that the odds of macroalbuminuria are 2.71 times higher in smokers than in nonsmokers.

Again, similar to the linear regression analysis, the logit of the outcome variable can be described by an equation including several independent or explanatory variables:

$$\text{logit } y = l_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots \beta_n x_n$$

The second aim of the study is to investigate whether the link between smoking and macroalbuminuria is confounded by age. We consider age as a potential confounder because it differed in smokers and nonsmokers (the exposure) as well as in patients with and without macroalbuminuria (the outcome), and because it cannot be considered as an effect of the exposure (i.e. age is not influenced by smoking).

To test whether the link between smoking and macroalbuminuria is independent of age, we introduce age into the multiple logistic regression analysis. Therefore, the logit equation becomes:

$$\text{logit of macroalbuminuria } (y) = \\ -2.87 + 0.97 \cdot \text{smoking } (0 = \text{no}; 1 = \text{yes}) + 0.06 \cdot \text{age (years)}$$

The odds ratio of macroalbuminuria (smokers vs. nonsmokers) corresponding to a regression coefficient of 0.97 is:

$$\text{odds ratio} = 2.7183^{0.97} = 2.63$$

Data adjustment of smoking for age did not materially modify the odds ratio of the relationship between smoking and the odds of macroalbuminuria (2.63 vs. 2.71). In other words, the link between smoking and the odds of macroalbuminuria is only slightly affected by data adjustment for age.

Table 2. Dependent variable: microalbuminuria (mg/24 h)

Explanatory variables	Units of increase	Regression coefficient, standard error and p value	
		crude analysis	age-adjusted analysis
Serum glucose	1 mg/dl	0.75 ± 0.19, p < 0.001	0.62 ± 0.20, p = 0.002
Age	1 year		0.55 ± 0.24, p = 0.02
Intercept		50.83 ± 19.29	38.95 ± 19.54

Standard error is directly provided in the computer output. For interested readers more details on standard error are reported elsewhere [8].

Number of Covariates Included into the Multiple Logistic Regression Analysis

The maximum number of variables that can be included into a multiple logistic regression model is dependent on the number of events rather than on the number of observations. A very simple rule is to include 1 variable for every 10 events into the multiple logistic regression model [7]. Thus, if we have a sample of 1,500 individuals who experienced 30 events during a given follow-up, the maximum number of covariates to include into the multiple logistic model should be 3.

Guidance on Presentation and Interpretation of Results

Linear Regression Analysis: Presentation of Results

In Example 1, the aim of the study was to investigate the relationship between serum glucose and microalbuminuria in a series of 98 hypertensive patients and whether this link is affected by the confounding effect of age. The results of the linear regression analysis are summarized in table 2, in which we reported the crude and the age-adjusted effect of serum glucose on microalbuminuria.

To provide the results of a linear regression analysis (either univariate or multiple), we have to indicate the dependent variable (microalbuminuria) as well as the list of explanatory variables (or independent variables; serum glucose and age), the unit of increase of each variable, the regression coefficient and its standard error and the p value. As discussed previously, it is preferable to report also the value of the intercept for predictive purposes. In fact, the intercept can be used, together with the regression coefficient, to predict the estimated value of 24-hour microalbuminuria for a given individual, of whom we know the serum glucose concentration.

Interpretation of Results

In table 2 we reported the crude and the age-adjusted effect on microalbuminuria of 1 mg/dl increase in serum glucose. As shown, a 1 mg/dl increase in serum glucose is associated with a 0.75 mg/24 h increase in microalbuminuria or, as a consequence of the linearity assumption underlying the linear regression, a 20 mg/dl increase in serum glucose determined a 15 mg/24 increase in microalbuminuria (0.75 · 20). Data adjustment for the confounding effect of age reduced the strength of the link between serum glucose and microalbuminuria (0.62 vs. 0.75) although serum glucose remained independently related to the outcome variable. Furthermore, age was directly related to microalbuminuria independently of serum glucose and a 1-year increase in age was associated with a 0.55 mg/24 h increase in microalbuminuria (that is, a 10-year increase in age is associated with a 5.5 mg/24 h increase in microalbuminuria).

As for the correlation coefficient (r), it is important to realize that this index [ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation)] is particularly well suited for predictive purposes. In our example, the correlation coefficient of the serum glucose to 24-hour microalbuminuria link is 0.33 (fig. 1). The square of the correlation coefficient (0.332 = 0.11, i.e. 11%) indicates that about 1/10 of the total variability in 24-hour microalbuminuria is explained by the variability in serum glucose. In summary, we can conclude that: (1) serum glucose is directly related to 24-hour microalbuminuria; (2) the increase in 24-hour microalbuminuria triggered by a 20 mg/dl increase in serum glucose is clinically relevant (+15 mg/24 h); (3) the relationship between serum glucose and 24-hour microalbuminuria is independent of age, and (4) given the relatively low correlation coefficient (r = 0.33), serum glucose cannot be used to predict the 24-hour microalbuminuria at an individual level in our study sample.

Table 3. Dependent variable: transition to macroalbuminuria

Explanatory variables	Units of increase	Odds ratio (95% confidence interval) and p value	
		crude analysis	age-adjusted analysis
Smoking	0 = no 1 = yes	2.71 (1.18–6.26), p = 0.019	2.65 (1.10–6.40), p = 0.03
Age	1 year		1.06 (1.02–1.10), p = 0.002
Crude analysis: intercept = -0.27 ± 0.26 . Age-adjusted analysis: intercept = -2.87 ± 0.92 .			

Logistic Regression Analysis: Presentation of Results

In Example 2 we described a study investigating the relationship between smoking (independent variable) and the transition from microalbuminuria to macroalbuminuria (dependent variable) in a series of 100 diabetic patients. In this instance, both the exposure (0 = non-smoker; 1 = smoker) and the dependent variable (0 = no transition to macroalbuminuria; 1 = transition to macroalbuminuria) are categorical variables (table 3).

The data presentation of a logistic regression analysis makes it necessary, as for the linear regression analysis, to indicate the dependent variable (transition to macroalbuminuria), the explanatory variables (smoking and age), the units of increase in the explanatory variables, the odds ratio, the 95% confidence interval and the p value. It is also preferable to indicate the value of the intercept (either crude or adjusted) for predictive purposes.

Interpretation of Results

In table 3 the odds ratio of smoking for macroalbuminuria is reported in both crude and age-adjusted terms. In the crude analysis the odds of macroalbuminuria are 2.72 times higher in smokers than in nonsmokers with a 95% confidence interval not including 1 (1.18–6.26). Data adjustment for age did not materially affect this effect (odds ratio: 2.65, 95% confidence interval: 1.10–6.40). Age resulted to be an independent variable related to macroalbuminuria and 1 year's increase in age was associated with a 6% increase (odds ratio: 1.06) in the odds of macroalbuminuria. If we want to know the increase in the odds of macroalbuminuria associated with 10 years' increase in age, a common mistake is to multiply the odds ratio (1.06) by 10. This approach is not valid because in the logistic regression analysis the underlying function is the log transformation of the odds and not a linear function. For this reason, to calculate the odds ratio of macroalbuminuria associated with 10 years' increase in age we have to perform the following steps:

(1) to go back to the regression coefficient by calculating the natural logarithm of the odds ratio:

$$\ln(1.06) = 0.058$$

(2) to multiple the regression coefficient by 10:

$$0.058 \times 10 = 0.58$$

(3) to calculate the exponential of this regression coefficient to have the odds ratio of macroalbuminuria associated with 10 years' increase in age:

$$2.7183^{0.58} = 1.79^a$$

Thus, a 10-year increase in age is associated with a 79% increase in the odds ratio of macroalbuminuria.

Guidance on Using Statistical Packages (Also by Including a Syntax)

Below, the syntaxes for linear and logistic regression analyses as provided for SPSS and SAS are listed.

SPSS

Linear Regression Analysis

Aim: to investigate the relationship between serum glucose and 24-hour microalbuminuria (24 h_{mU}) by considering the confounding effect of age:

- including regression coefficients, standard errors, p value and residuals plotting
- ```

REGRESSION
/DESCRIPTIVES MEAN STDDEV CORR SIG N
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA = PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT 24h_mU
/METHOD = ENTER glucose age
/RESIDUALS HIST(ZRESID).
```

<sup>a</sup> This value equals 1.06 (the odds ratio of macroalbuminuria associated with 1 year's increase in age) to the tenth power ( $1.06^{10} = 1.79$ ).

## Logistic Regression Analysis

Aim: to investigate the relationship between smoking (0 = no; 1 = yes) and transition from microalbuminuria to macroalbuminuria (event) by considering the confounding effect of age:

- including regression coefficients, standard errors, odds ratio, 95% confidence interval and p value
- ```
LOGISTIC REGRESSION VAR = event  
/METHOD = ENTER smoking age  
/PRINT = CI(95)  
/CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5).
```

SAS

Linear Regression Analysis

The name of the file is 'example1' and it is resident in 'c:\sasreg'.

To upload the dataset in SAS

```
set 'c:\sasreg\example1.sas7bdat';  
run;
```

To perform the analysis

```
proc reg data = example1;  
model 24h_mU = glucose age;  
run;
```

Logistic Regression Analysis

To upload the dataset in SAS

```
set 'c:\sasreg\example2.sas7bdat';  
run;
```

To perform the analysis

```
proc logistic data = example2 descending;  
model event = smoking age;  
run;
```

Conclusions

The regression analysis describes the dependence of the outcome variable (or dependent variable) from one or more explanatory variables (or independent variables) by providing a mathematical model allowing the prediction of the outcome variable when we know the value of the predictor variable. The linear relationship between exposure (either continuous or categorical) and a continuous outcome can be assessed by using linear regression analysis. By contrast, if the outcome is dichotomic (e.g. dead/alive or presence/absence of a given disease) and the exposure is either continuous or categorical, their relationship can be tested by logistic regression analysis.

Acknowledgment

This study is part of the SysKID project which is supported through the European Union's FP7, grant agreement No. HEALTH-F2-2009-241544.

References

- 1 Tripepi G, Jager KJ, Dekker FW, Zoccali C: Stratification for confounding. 1. The Mantel-Haenszel formula. *Nephron Clin Pract* 2010;116:c317–c321.
- 2 Tripepi G, Jager KJ, Dekker FW, Zoccali C: Stratification for confounding. 2. Direct and indirect standardization. *Nephron Clin Pract* 2010;116:c322–c325.
- 3 van Stralen KJ, Dekker FW, Zoccali C, Jager KJ: Confounding. *Nephron Clin Pract* 2010;116:c143–c147.
- 4 Armitage P, Berry G: *Statistical Methods in Medical Research*, ed 3. London, Blackwell, 1994.
- 5 Kleinbaum DG, Kupper LL, Muller KE, Nizam A: *Applied Regression Analysis and Other Multivariable Methods*. Pacific Grove, Duxbury Press, 1998, pp 43–47, 251–252, 389–390, 639–655.
- 6 Tripepi G, Jager KJ, Dekker FW, Zoccali C: Measures of effect in epidemiological research. *Nephron Clin Pract* 2010;115:c91–c93.
- 7 Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR: A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996;49:1373–1379.
- 8 Kleinbaum DG, Sullivan KM, Barker ND: *A Pocket Guide to Epidemiology*. New York, Springer, 2007, pp 198, 204–208, 216.