Linear and logistic regression analysis

G Tripepi[1], KJ Jager[2], FW Dekker[2,3] and C Zoccali[1]

[1]*CNR-IBIM, Clinical Epidemiology and Physiopathology of Renal Diseases and Hypertension of Reggio Calabria, Reggio Calabria, Italy;* [2]*ERA–EDTA Registry, Department of Medical Informatics, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands and* [3]*Department of Clinical Epidemiology, Leiden University Medical Centre, Leiden, The Netherlands*

In previous articles of this series, we focused on relative risks and odds ratios as measures of effect to assess the relationship between exposure to risk factors and clinical outcomes and on control for confounding. In randomized clinical trials, the random allocation of patients is hoped to produce groups similar with respect to risk factors. In observational studies, exposed and unexposed individuals may differ not only for the presence of the risk factor being tested but also for a series of other factors that are potentially related to the study outcome, thus making 'confounding' very likely. One of the most important uses of multivariate modeling is precisely that 'of controlling for confounding' to let emerge the effect of the risk factor of interest on the study outcome. In this paper, we will discuss linear regression analysis for the examination of continuous outcome data and logistic regression analysis for the study of categorical outcome data. Furthermore, we focus on the most important application of multiple linear and logistic regression analyses.

Kidney International (2008) **73,** 806–810; doi:10.1038/sj.ki.5002787; published online 16 January 2008

KEYWORDS: epidemiology; linear regression analysis; logistic regression analysis; confounding

Correspondence: *G Tripepi, CNR-IBIM, Istituto di Biomedicina, Epidemiologia Clinica e Fisiopatologia, delle Malattie Renali e dell'Ipertensione Arteriosa, c/o Euroline di Ascrizzi Vincenzo, Via Vallone Petrara n. 55/57, Reggio Calabria 89124, Italy. E-mail: gtripepi@ibim.cnr.it*

Received 5 September 2007; revised 16 October 2007; accepted 6 November 2007; published online 16 January 2008

LINEAR CORRELATION, AND SIMPLE AND MULTIPLE LINEAR REGRESSION ANALYSES

Linear correlation analysis

Correlation and regression analyses are based on identical calculations but address different questions. Correlation analysis investigates the *degree of association* between two continuous variables, that is, it defines how much a given relationship is fitted by a straight line. In correlation analysis, the investigator is simply interested in estimating the strength of linear association between two variables. In general, this analysis is applied to estimate the degree of association between two variables when there is no sufficient knowledge to identify which of the two is responsible for the variability in the other variable or when this information is irrelevant to the question being asked. Regression analysis instead is used to describe *the linear dependence* of the outcome variable (or dependent variable) from one (or more) predictor variable (or independent variable).

In a recent paper,[1] the relationship between serum albumin and free triiodothyronine (plasma levels of fT3) was investigated in a sample of 41 patients on chronic ambulatory peritoneal dialysis.

There is experimental evidence that malnutrition and inflammation impair thyroid function and therefore the investigators decided to identify plasma fT3 as the outcome variable (or dependent variable) and serum albumin (a direct marker of malnutrition and an inverse marker of inflammation as well) as the predictor variable (or independent variable). In regression analysis, the predictor variable is always plotted on the horizontal axis (the X scale) and the outcome variable on the vertical axis (the Y scale). Each dot in the graph represents an individual and it is identified by a pair of values: the value of albumin and the corresponding value of fT3. The scatter plot in Figure 1 (left panel) shows that plasma fT3 increases in parallel with serum albumin and vice versa, suggesting a linear relationship between the two variables. In our example, the correlation coefficient of the albumin–fT3 link is 0.52. The square of the correlation coefficient ($0.52^2 = 0.27$, that is, 27%) indicates that about 1/4 of the total variability in plasma fT3 is explained by concomitant variability in serum albumin. Linear association does not demand the two variables changing in the same direction. Indeed, two variables may be linearly related also when they change in opposite directions (the fT3–age link plotted in Figure 2) and in such a situation the correlation coefficient is negative ($r = -0.61$, $P < 0.001$).

Linear regression, intercept, regression coefficient, and residuals

The linear dependence between serum albumin and plasma fT3 can be assessed by calculating the increase in plasma fT3 for each unitary increase in serum albumin. This information can be obtained by using the regression line, that is, a line that 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, β_0 is the intercept, β_1 is the regression coefficient, and x is a given value of the independent or predictor variable.

The intercept (β_0) is the theoretical value of Y when X equals zero (Figure 1, left panel). The regression coefficient (β_1) is the estimated increase in the dependent variable (Y) per one unit increase in the independent variable (X) or the slope of the regression line (that is, the tangent of the angle between the regression line and the X axis) (Figure 1, left panel). The method used to estimate the intercept and the regression coefficient is the least squares method. This method consists of finding the parameters (β_0 and β_1) that minimize the sum of the squares of the vertical deviations of observed data points and the predicted values in the regression line (see vertical pointed lines in Figure 1, right panel). These deviations are called *residuals*. The least squares method is described in full detail elsewhere.[2]

The mathematical equation for the regression line of the fT3–albumin link in our sample, as provided by the computer output, is

$$\text{estimated fT3} = -1.84 + 1.36 \times \text{albumin (g per 100ml}^{-1})$$

A regression coefficient of 1.36 means that for each 1 g per 100 ml change in serum albumin, there is a corresponding change of 1.36 pg ml^{-1} in plasma fT3 (for example, for 2 g per 100 ml decrease in serum albumin, there is an average decrease of 2.72 pg ml^{-1} (that is, 1.36 × 2) in plasma fT3). A positive regression coefficient indicates a direct relationship between risk factor and outcome variable and a negative regression coefficient indicates an inverse one. A regression coefficient close to zero indicates no association. The value of the intercept (-1.84 pg ml^{-1}) corresponds to the estimated fT3 level when albumin is zero (Figure 1, left panel). Clearly, a negative value of fT3 (-1.84 pg ml^{-1}) and a serum albumin of zero are purely theoretical values. The intercept is useful because it can be applied, together with the regression coefficient, to predict the estimated value of plasma fT3 for a given individual of which we know the corresponding serum albumin concentration. For example, the estimated value of plasma fT3 for an individual having a serum albumin of

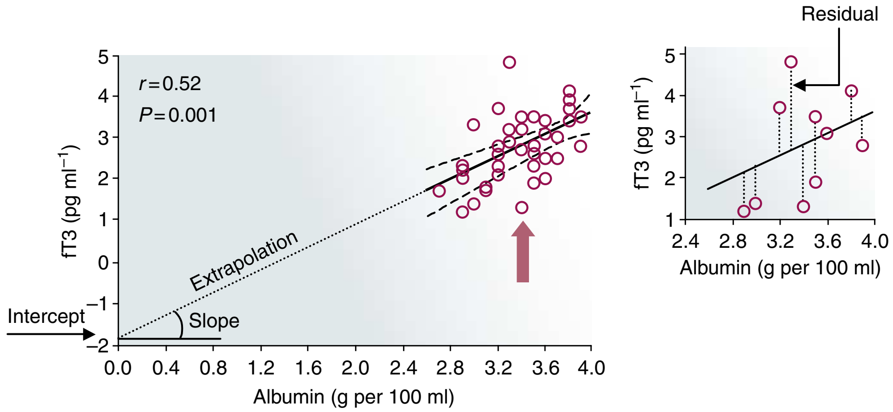

Figure 1 | Relationship between serum albumin and plasma fT3 (left panel) in 41 patients on chronic ambulatory peritoneal dialysis.[1] In the right panel, the concept of 'residual' is described graphically as the distance (vertical pointed lines) between each observed value and the regression line (see text for more details).

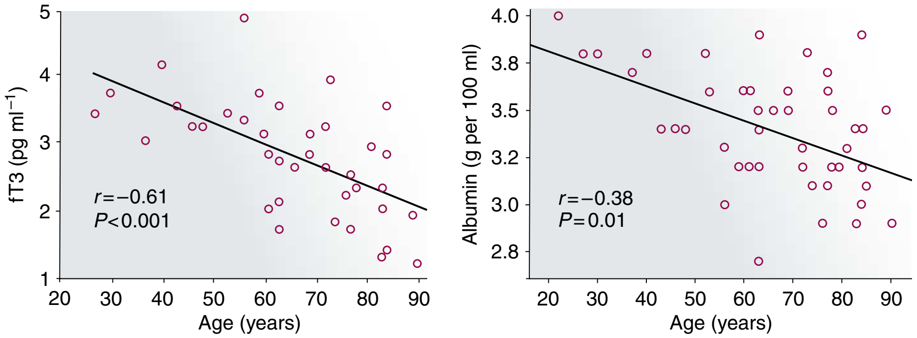

Figure 2 | Relationship between age with plasma fT3 and serum albumin in 41 patients on chronic ambulatory peritoneal dialysis.[1] Data are Pearson correlation coefficient and *P*-value.

3.4 g per 100 ml (see dot indicated by the arrow in Figure 1, left panel) can be easily calculated by resolving the equation

$$\text{estimated fT3} = -1.84 + 1.36 \times 3.4 = 2.78 \text{pgml}^{-1}$$

Thus, by using the regression line constructed in our sample, we predict a plasma fT3 of 2.78 pg ml^{-1} for an individual with a serum albumin of 3.4 g per 100 ml. For this individual, the residual is calculated as the difference between the observed (1.30 pg ml^{-1}) and the estimated value of serum albumin (2.78 pg ml^{-1}), which is -1.48 pg ml^{-1}.

By repeating this calculation for all observed and predicted values, we obtain the distribution of residuals. The analysis of residuals is of particular relevance for the 'diagnostics' of linear regression analysis. Regression diagnostics rests on three assumptions: (1) that to each value of the independent variable corresponds a set of normally distributed values of the dependent variable; (2) that the standard deviation of this set of values is the same for each value of the independent variable; and (3) that the relationship between the two variables is linear. If all these assumptions are true, the residuals should be normally distributed. In our instance, the residuals of the fT3–albumin link have an approximately normal distribution (not shown), implying that the data distribution in the sample meets all the above-mentioned criteria.

95% confidence interval of the regression line

The regression line of the fT3–albumin link we fitted in our sample is an estimate of the 'true' regression line, that is, of the regression line of fT3–albumin link in the theoretical population that includes all dialysis patients worldwide. Therefore, we need to compute the degree of uncertainty of our estimate by calculating the 95% confidence interval (or prediction interval) of the regression line (see dotted lines in Figure 1, left panel). The concept of the confidence interval for the regression line can be explained as follows: if we draw 100 samples of the same size as the study sample ($n = 41$) from the chronic ambulatory peritoneal dialysis population and calculate for each sample the regression line of the fT3–albumin link, we obtain a family of 100 (slightly different) regression lines. The 95% confidence interval is the interval that includes 95% of the regression lines of these 100 study samples. In our instance, the 95% confidence interval is fairly narrow, indicating that the linear model provides an adequate data fitting of the fT3–albumin link.

Multiple linear regression analysis

In another paper of this series,[3] it was discussed that 'confounding' may disturb the interpretations of the effect of an exposure on outcome. In that paper, we showed that confounding can be prevented by randomization, restriction, or matching, that is, by an appropriate study design. Confounding may also be dealt with in the analytical phase of the study by stratification or multiple linear regression analysis that nicely serves this scope. Multiple linear regression analysis allows estimation of the linear effect of a given independent variable (for example, x_1) on a given dependent or outcome variable (y) after controlling for the confounding effect of other variables (or covariates) (for example, x_2, x_3, …, x_n). The corresponding multiple linear regression model is

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

where $E(y)$ is the estimated or predicted value of Y, β_0 is the intercept (that is, the value of Y when x_1, x_2, and x_3 are zero), and β_1, β_2, β_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 link between plasma fT3 and serum albumin in 41 patients on chronic ambulatory peritoneal dialysis and found that the two variables were strongly inter-related. Now, we analyze the effect of serum albumin on plasma fT3 by adjusting for the confounding effect of age, a variable that was linearly related with plasma fT3 ($r = -0.61$, $P < 0.001$) and serum albumin ($r = -0.38$, $P = 0.01$) (Figure 2). We consider age as a potential confounder because it meets criteria set for the definition of confounder.[3] In fact, age influences both plasma fT3 (the outcome variable) and serum albumin (the predictor variable); cannot be considered as an effect of exposure (albumin as an indicator of malnutrition/inflammation); and we assume that age is not in the causal pathway between exposure (serum albumin) and outcome (plasma fT3). After introducing age into the multiple linear model, the regression line provided by the computer output is

$$\text{Estimated fT3} = 1.41 + 0.87 \times \text{albumin (g per 100ml)} - 0.024 \times \text{age (years)}$$

A 0.87 regression coefficient for serum albumin means that for each 1 g per 100 ml change in this variable, there is a 0.87 pg ml^{-1} change in plasma fT3 and this estimate is adjusted for the confounding effect of age. Comparing the adjusted effect (0.87) and the unadjusted effect reported above (1.36), we see that indeed age was a confounder here as adjustment for age changed the effect of albumin on fT3. The results of the multiple linear regression analysis are summarized in Table 1.

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 every 10 observations.[4] Thus, if we are to construct a model based on 10 variables, the general rule demands a sample size of 100 individuals.

SIMPLE AND MULTIPLE LOGISTIC REGRESSION ANALYSIS

Simple logistic regression analysis

Linear regression analysis demands that the dependent variable is 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) (that is, 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.

A recent study in a series of 500 patients with essential hypertension[5] investigated the relationship between systemic endothelial dysfunction (as defined on the basis of the maximal vasodilatory response to the infusion of acetylcholine in the forearm) and the risk of chronic kidney disease (CKD, defined as a glomerular filtration rate <60 ml per min per 1.73 m^2). Hypertensive patients were classified as having ($n = 73$) or not having ($n = 427$) CKD and divided into two categories on the basis of the maximal vasodilatory response to acetylcholine (ACh) in the forearm ($<400\%$: endothelial dysfunction; $\geqslant 400\%$: normal endothelial function). See Table 2 for the proportion of individuals with CKD in these two categories.

The proportion of individuals with CKD in patients with endothelial dysfunction was about three times (0.17, that is, 17.0%) that in those with normal endothelial function (0.063, that is, 6.3%) (Table 2).

Odds, odds ratio, and logit

The odds of CKD (third column) were calculated by the formula

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

In the group of individuals with a response to ACh $<400\%$, the odds of CKD were

$$\text{odds} = 0.170/(1-0.170) = 0.205$$

In the group of individuals with a response to ACh $\geqslant 400\%$, the odds of CKD were

$$\text{odds} = 0.063/(1-0.063) = 0.067$$

Thus, the odds ratio (OR) of CKD between patients with and without endothelial dysfunction will be the ratio between the two odds:

$$\text{OR} = 0.205/0.067 = 3.06$$

Table 1 | Multiple linear regression analysis of plasma fT3

Multiple R=0.68, $P<0.001$		
Covariates (units of measure)	**Regression coefficients**	**P**
Serum albumin (g per 100 ml)	0.87	0.01
Age (years)	−0.024	0.001
Intercept (β_0)	1.41	0.31

fT3, triiodothyronine.

The subsequent step was to make the logit (or logistic) transformation of the odds of CKD (see Table 2, last column). The logit is the natural logarithm (ln) of the odds.

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

where p is the proportion of individuals with CKD in each category of maximal response to ACh. For example, the logit transformation of the odds of CKD in the group of individuals with a response to ACh $<400\%$ is

$$\text{logit} = \ln(0.205) = -1.58$$

As in linear regression analysis, in logistic regression analysis also the outcome (dependent) variable is described by a simple equation:

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

To be able to interpret this simple equation, both sides of the equal to sign could be raised to the power e = 2.7183. It can be shown that reworking this equation results in a nice interpretation: e^{b1} is the OR of one unit increase in x. In this analysis, the intercept (β_0) is the value of the natural logarithm of the odds of CKD when endothelial function equals zero and the regression coefficient (β_1) is the logarithm of the odds of CKD in patients with endothelial dysfunction. In the logistic regression analysis, the regression coefficients are calculated by using the maximum likelihood method, that is, a method that by an iterative calculation routine identifies the regression coefficients that maximize the probability of the observed data.[7] The regression coefficients are directly provided by the print-out of the statistical software. To estimate the increase in the risk of CKD in patients with endothelial dysfunction as compared to that of patients with normal endothelial function, the authors made the inverse operation of logit transformation, that is, calculated the antilogarithm of the regression coefficient. In other words, they computed the OR by exponentiating the base of the natural logarithm (e = 2.1783) to the regression coefficient (β_1): 2.7183^{β_1}. Therefore, the OR corresponding to a regression coefficient of 1.118 (see Table 3) is

$$\text{OR ratio} = 2.7183^{1.118} = 3.06$$

Of note, the OR calculated by univariate logistic regression analysis is identical to that calculated by starting with the odds of CKD in patients with and without endothelial dysfunction (see above).

The study indicated that patients with endothelial dysfunction (hemodynamic response to ACh $<400\%$) had

Table 2 | Relationship between the maximal vasodilatory response to ACh and the risk of CKD

Maximal vasodilatory response to ACh (%)	Proportion of individuals with CKD (p)	Odds of CKD: $p/(1-p)$	Logit (ln odds)
$\geqslant$400 (n=111) (normal endothelial function)	0.063	0.067	−2.70
$<$400 (n=389) (endothelial dysfunction)	0.170	0.205	−1.58

ACh, acetylcholine; CKD, chronic kidney disease.
The concept of odds is described in a previous article of this series.[6] A higher vasodilatory response to ACh denotes better endothelial function.

Table 3 | Simple logistic regression analysis of CKD (GFR < 60 ml per min per 1.73 m^2)

	Units of increase (ordered group)	Regression coefficient (β_1)	OR (95% confidence interval)	*P*
Maximal vasodilatory response to ACh	**0** (⩾400%) **1** (<400%)	1.118	1.00 (reference group) 3.06 (1.35–6.82)	0.007
Intercept (β_0)=−2.70				<0.001

ACh, acetylcholine; CKD, chronic kidney disease; GFR, glomerular filtration rate; OR, odds ratio.
A higher vasodilatory response to ACh denotes better endothelial function.

Table 4 | Multiple logistic regression analysis of CKD

	Units of increase (ordered group)	Regression coefficient (β_1)	OR (95% confidence interval)	*P*
Maximal vasodilatory response to Ach	**0** (⩾400%) **1** (<400%)	0.974	1.00 (reference group) 2.64 (1.17–6.00)	0.02
Age	1 year	0.047	1.05 (1.02–1.07)	<0.001
Intercept (β_0)=−4.91				<0.001

ACh, acetylcholine; CKD, chronic kidney disease; OR, odds ratio.
A higher vasodilatory response to ACh denotes better endothelial function.

an OR of CKD that was about three times that in those with normal endothelial function (reference category: OR = 1). This finding is of clinical relevance because the 95% confidence interval does not include 1 (see Table 3).

Multiple logistic regression analysis

Again in close similarity with linear regression analysis, the logit of the outcome variable can be described by an equation including several independent (or predictor) variables:

$$\text{logit}\, y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n$$

To further elaborate on the same example, we now wonder whether the link between endothelial function and CKD is confounded by age. We consider once again age as a potential confounder because it affects the risk of CKD (the outcome variable) and the risk of endothelial dysfunction (the predictor variable) and because it cannot be considered as an effect of the exposure (that is, age is not influenced by endothelial function).

To test whether the link between endothelial function and the risk of CKD is independent of age, we introduce age into the multiple logistic regression analysis. The results of this analysis are summarized in Table 4.

Adjustment of maximal vasodilatory response to ACh for age did not materially modify the OR of the relationship between endothelial function and the risk of CKD (2.64 vs 3.06). In other words, the link between endothelial function and the risk of CKD was only slightly affected by age.

Number of covariates in the multiple logistic regression analysis

The maximum number of covariates that can be included in a multiple logistic regression model is strictly dependent on the number of events rather than on the number of observations. A simple rule is to include in the multiple logistic regression model 1 covariate every 10 events.[8] For example, if we have a sample of 1000 individuals who experienced 20 events during a given follow-up, the maximum number of covariates to include in the multiple logistic model should be 2.

CONCLUSION

Linear and logistic regression analyses are important statistical tools for assessing relationships between exposure and outcome and for controlling confounding in epidemiological studies. Here we focused on their use in etiological research.

The validity of any conclusion drawn by using these methods is critically dependent on the ascertainment of a series of assumptions. The lack of a rigorous validation of these conditions may lead to flawed data analyses and invalid results.

REFERENCES

1. Enia G, Panuccio V, Cutrupi S *et al.* Subclinical hypothyroidism is linked to micro-inflammation and predicts death in continuous ambulatory peritoneal dialysis. *Nephrol Dial Transplant* 2007; **22**: 538–544.
2. Armitage P, Berry G. *Statistical Methods in Medical Research*, 3rd edn. Blackwell: London, England, 1994.
3. Jager KJ, Zoccali C, MacLeod A, Dekker FW. Confounding: what it is and how to deal with it. *Kidney Int* 2007; October 31 [E-pub ahead of print].
4. Kleinbaum DG, Kupper LL, Muller KE, Nizam A. *Applied Regression Analysis and Other Multivariable Methods*. Duxbury Press: Pacific Grove, 1998, pp 389–390.
5. Perticone F, Maio R, Tripepi G, Zoccali C. Endothelial dysfunction and mild renal insufficiency in essential hypertension. *Circulation* 2004; **110**: 821–825.
6. Tripepi G, Jager KJ, Dekker FW *et al.* Measures of effect: relative risks, odds ratios, risk difference, and 'number needed to treat'. *Kidney Int* 2007; **72**: 789–791.
7. Kleinbaum DG, Kupper LL, Muller KE, Nizam A. *Applied Regression Analysis and Other Multivariable Methods*. Duxbury Press: Pacific Grove, 1998, pp 639–655.
8. Peduzzi P, Concato J, Kemper E *et al.* A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996; **49**: 1373–1379.