—Review—

Applications of Statistics to Medical Science, III
Correlation and Regression

Hiroshi Watanabe

Department of Mathematics, Nippon Medical School

Abstract

In this third part of a series surveying medical statistics, the concepts of correlation and regression are reviewed. In particular, methods of linear regression and logistic regression are discussed. Arguments related to survival analysis will be made in a subsequent paper.
(J Nippon Med Sch 2012; 79: 115–120)

Key words: statistical test, statistical inference, regression, correlation

1. Introduction

In the previous works[1,2], we studied the conceptual framework of statistical analysis and actual procedures for general use. The present work is devoted to regression methods: linear regression and logistic regression.

Textbooks[3,4] are rich sources of information on medical statistics. For details of logistic regression, see[5].

2. Linear Regression

In this section, we consider linear regression.

Suppose that we observe n subjects and measure two quantities X and Y for each subject. Let (x_1, y_1), (x_2, y_2), ⋯, (x_n, y_n) be obtained data, that is, x_j and y_j denote the values of X and Y, respectively, for the jth subject.

If we draw a scatter plot of the data, we can roughly grasp relationship between X and Y. Furthermore, we may see that the quantity Y is approximately expressed by some function $f(X)$. The task to find such a function $f(X)$ is called regression. If a suitable regression is found, it means that behavior of Y is at least partially explained by that of X. In particular, regression by means of a linear function

$$f(X)=a+bX \tag{1}$$

is referred to as linear regression.

Regression Coefficients

Given experimental data (x_1, y_1), (x_2, y_2), ⋯, (x_n, y_n), the optimal coefficients a and b in the right hand of (1) are determined by means of the *method of least squares*. The results are written as

$$a=\bar{y}-r_{xy}\frac{s_y}{s_x}\bar{x}=\bar{y}-b\bar{x}, \tag{2}$$

$$b=r_{xy}\frac{s_y}{s_x}. \tag{3}$$

We have denoted sample means by $\bar{x}$ and $\bar{y}$, standard deviations (not unbiased) by s_x and s_y, and **Pearson's correlation coefficient** by r_{xy}, i.e.,

Correspondence to Hiroshi Watanabe, Department of Mathematics, Nippon Medical School, 2–297–2 Kosugi-cho, Nakahara-ku, Kawasaki, Kanagawa 211–0063, Japan
E-mail: watmath@nms.ac.jp
Journal Website (http://www.nms.ac.jp/jnms/)

$$r_{xy}=\frac{c_{xy}}{s_x s_y}, \tag{4}$$

where c_{xy} stands for a covariance

$$c_{xy}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}). \tag{5}$$

Interval Estimation of Regression Coefficients

The regression coefficients a and b given by (2) and (3) are considered to be point estimates of the corresponding population parameters a_* and b_*. Relation between (a, b) and (a_*, b_*) is analyzed by means of the following probabilistic model.

We introduce a random variable ε_i associated with the ith subject and assume the relation between x_i and y_i as follows:

$$y_i=a_*+b_*x_i+\varepsilon_i,\ i=1, 2, \cdots, n, \tag{6}$$

where a_* and b_* are unknown population parameters. This means that the value y_i of Y has a deterministic part $a_* + b_* x_i$ and a stochastic part ε_i. Here, ε_1, ε_2, $\cdots$, ε_n are assumed to be independent random variables that obey the same normal distribution with mean 0 and variance σ^2. These assumptions are equivalent with the statement that y_1, y_2, $\cdots$, y_n obey independent normal distributions with mean $a_* + b_* x_i$ and variance σ^2 for i=1, 2, $\cdots$, n, respectively. The sample values x_1, x_2, $\cdots$, x_n of X are regarded as given constants, hence we need not assume a probability law for X.

We have to infer the values a_*, b_* and σ^2 from a sample. The estimators for a_* and b_* are given by (2) and (3), and σ^2 is estimated by $SS_{residual}/(n-2)$, where $SS_{residual}$, a *residual sum of squares*, is defined by

$$SS_{residual}=\sum_{i=1}^{n}(a+bx_i-y_i)^2. \tag{7}$$

Here, a and b are given by (2) and (3). Furthermore, squared standard errors of a and b are given by

$$SE(a)^2=\frac{SS_{residual}}{(n-2)n}\left(1+\frac{\bar{x}^2}{s_x^2}\right), \tag{8}$$

$$SE(b)^2=\frac{SS_{residual}}{(n-2)ns_x^2}, \tag{9}$$

respectively. Using the above formulae, we can make interval estimations of a and b. Especially for b, a statistic

$$t=\frac{b-b_*}{SE(b)} \tag{10}$$

obeys a t distribution with $n-2$ degrees of freedom, hence the confidence interval of b with confidence level $1-\alpha$ is written as

$$b-t_{n-2}(1-\alpha/2)SE(b)<b_*<b+t_{n-2}(1-\alpha/2)SE(b), \tag{11}$$

where $t_{n-2}(1-\alpha/2)$ stands for the $(1-\alpha/2)\times 100$th percentile of a t distribution with $n-2$ degrees of freedom.

Statistical Test

We can assess goodness of fit for the linear regression described above by performing a statistical test for a null hypothesis $H_0 : b_*=0$. The hypothesis H_0 means that behavior of Y is not at all explained by that of X. Then, rejection of H_0 means that behavior of Y is to some extent explained by that of X as (1).

A statistical test for H_0 is done by means of (11), that is, we accept H_0 if and only if $b_*=0$ satisfies (11). We can also perform the following test for H_0 based on the principle of ANOVA. Define a *regression sum of squares* $SS_{regression}$ by

$$SS_{regression}=\sum_{i=1}^{n}(a+bx_i-\bar{y})^2, \tag{12}$$

where a and b are given by (2) and (3). Then, under the null hypothesis H_0, a statistic

$$F=\frac{SS_{regression}}{\frac{1}{n-2}SS_{residual}} \tag{13}$$

obeys an F distribution with 1 and $n-2$ degrees of freedom. Therefore, H_0 is tested with significance level α by the rule:

Reject H_0 if only if $F>F^1_{n-2}(\alpha)$,

where $F^1_{n-2}(\alpha)$ denotes the $\alpha\times 100$th percentile of the F distribution with 1 and $n-2$ degrees of freedom.

Remarks. 1) The above two tests by (10) and by (13) are in fact equivalent because we can show $t^2=F$ under the assumption $b_*=0$.

2) Effects of two or more exposures on an outcome can be analyzed by appealing for **multivariate analysis**, in which the linear function (1) is replaced by

$$f(X_1, X_2, \cdots, X_k)=a+b_1X_1+b_2X_2+\cdots+b_kX_k. \tag{14}$$

The optimal coefficients a, b_1, b_2, $\cdots$, b_k are determined

by the method of least squares and written in terms of linear algebra.

3) We may want to estimate a population correlation coefficient r_{xy*} instead of the population regression coefficient b_*. A null hypothesis $H_0 : r_{xy*} = 0$ is tested on the basis of the fact that, if X and Y obey a (two dimensional) normal distribution and if $r_{xy*} = 0$, the statistic

$$t = \frac{\sqrt{n-2}\, r_{xy}}{\sqrt{1-r_{xy}^2}} \tag{15}$$

obeys a t distribution with $n - 2$ degrees of freedom, Interval estimation of r_{xy*} is also possible by means of z transformation.

4) We may sometimes be interested in *concordance* of X and Y instead of their correlation. For example, when two persons rate some characteristics, e.g., stages of a disease, that cannot be measured objectively, we may want to assess reproducibility (reliability) of the ratings. This problem would seem to be dealt with as a problem of linear regression by the function $Y=X$ and solved by testing the hypothesis "$a_* = 0$ and $b_* = 1$" in (6). However, this approach is inappropriate by several reasons. We should use the **concordance correlation coefficient** (for pairs) or the **overall concordance correlation coefficient** (for general cases). The **intraclass correlation coefficient** is also available. For categorical data, **κ coefficient** should be used.

3. Logistic Regression

We sometimes encounter problems to which the method of linear regression by the function (1) or (14) is not successful because of the simplicity of the function. Let p be probability of an outcome O and let X be a biometric quantity that measures a certain exposure. To judge whether the exposure is a risk factor of O or not, we assume that p is determined by X and write

$$p = f(X), \tag{16}$$

where the function $f(X)$ is suitably chosen according to a given sample. If $f(X)$ turns out to be a constant independent of X, we conclude that the exposure is not a risk factor. For this purpose, the function (1) is inappropriate, because values of the function $f(X) = a + bX$ are not restricted in the interval $0 \leq f(X) \leq 1$. Logistic regression is a solution that is frequently chosen for this kind of problems.

Table 1 Example data for logistic regression. Values of an exposure variable X and of an outcome variable Y are shown for ten subjects. If $Y=1$ (or $Y=0$), it means that the outcome is observed (or not observed, respectively)

Subject	X (exposure)	Y (outcome)
1	2.0	1
2	2.2	0
3	2.3	1
4	2.7	0
5	2.8	0
6	2.9	0
7	3.0	0
8	3.1	0
9	3.2	0
10	3.3	0

Example

Suppose that we observe ten subjects and estimate probability p of a certain outcome O. If two subjects have the outcome O and the remaining eight do not, it is reasonable to make an estimate $p = 0.2$. This estimation may be altered if we have additional data on a biometric quantity X for the ten subjects as in **Table 1**, because the probability p can depend on X. Our interest is to find a function $f(X)$ in (16) and judge whether X is a risk factor of O or not.

Let x_i and y_i be values of X and Y, respectively, for the ith subject. If $y_i=1$ (or $=0$), the probability $f(x_i)$ of O for the ith subject should be estimated to be large (or small), i.e., near to 1 (or 0, resp.). In short, we expect that $f(x_i)$ may be near to y_i, i.e.,

$$y_i \fallingdotseq f(x_i),\ i = 1, 2, \cdots, 10. \tag{17}$$

As is shown in **Figure 1**, the result of regression by means of a linear function $f(X) = a + bX$ is not satisfactory because deviation of the plot from a line is large, whereas regression by a certain curve seems to be better. This curve corresponds to the function

$$f(X) = \frac{e^{a+bX}}{1+e^{a+bX}}, \tag{18}$$

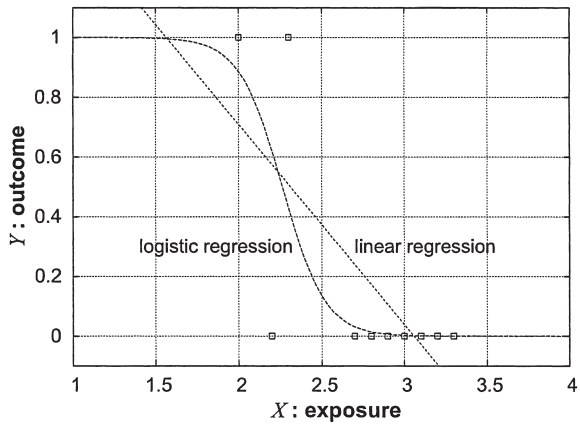

Fig. 1 Comparison of linear regression and logistic regression. The data given by Table 1 are drawn by small squares on an XY plane. Logistic regression is obviously better than linear regression.

with

$$a = 17.718,\ b = -7.825. \tag{19}$$

where e=2.718281828… is a mathematical constant. Note that the positivity $e^{a+bX}>0$ implies the inequality $0<f(X)<1$.

Logistic Regression Model

Regression by a function in the form of (18) is called **logistic regression**. Since $f(X)$ gives a probability, the quantity

$$e^{a+bX}=\frac{f(X)}{1-f(X)} \tag{20}$$

gives an odds. If $b>0$ (or <0), the odds e^{a+bX} is increasing (or decreasing) with respect to X, hence the model describes a phenomenon that is likely to be observed when X is large (or small, respectively). If $b=0$, the odds is independent of X, that is, X is not a risk factor of the outcome.

The optimal values of a and b are determined by means of the *method of maximum likelihood*, which is a generalization of the method of least squares. The method of least squares itself cannot be applied to this problem because it is justified under the assumption that Y is normally distributed. For the data given in **Table 1**, Y is categorical and y_i is either 0 or 1 for each i, hence the distribution of Y is far from normal and we cannot assume a model as (6). This is the true reason for avoiding the method of least squares in logistic regression.

We cannot write explicit formulae (as (2) and (3)) for the optimal values of a and b determined by the method of maximum likelihood. Instead, we numerically obtain the optimal values by using software packages as SAS and SPSS. The values shown in (19) were obtained by SPSS.

Interval Estimation of Regression Coefficients

The regression coefficients a and b chosen as above are considered to be estimates of the corresponding population parameters a_* and b_*, respectively. In this notation, we should write the logistic regression model as follows:

$$p=\frac{e^{a_*+b_*X}}{1+e^{a_*+b_*X}}. \tag{21}$$

Software packages as SAS and SPSS usually output, in addition to values of a and b, their standard errors SE(a) and SE(b). In view of these outputs, we can make interval estimation. In particular, because the statistic

$$u=\frac{b-b_*}{\mathrm{SE}(b)} \tag{22}$$

approximately obeys the standard normal distribution, a confidence interval for b_* with confidence level $1-\alpha$ is

$$b-z(1-\alpha/2)\,\mathrm{SE}(b)<b_*<b+z(1-\alpha/2)\,\mathrm{SE}(b), \tag{23}$$

where $z(1-\alpha/2)$ denotes the $(1-\alpha/2)\times 100$th percentile of the standard normal distribution.

For the data given in **Table 1**, we have b=−7.825 and SE(b)=6.367 according to SPSS, from which we obtain the following confidence interval with confidence level of 5%:

$$-16.152<b_*<4.654. \tag{24}$$

Odds Ratio

In an epidemiological context, we may want to estimate an odds ratio. In the logistic regression model (21), $e^{a_*+b_*X}$ gives an odds. If X is a categorical variable whose value is 0 or 1, the ratio r_* of odds $e^{a_*+b_*X}$ for X=1 and for X=0 is given by

$$r_*=\frac{e^{a_*+b_*1}}{e^{a_*+b_*0}}=e^{b_*}. \tag{25}$$

In view of the above formula for the odds ratio r_*,

we see that r_* has a point estimator e^b and an interval estimate has the form

$$e^{b-} < r_* < e^{b+}, \tag{26}$$

where b is the point estimator of b_* and $b_\pm$ stand for the limits $b \pm z(1-\alpha/2)\mathrm{SE}(b)$ of the confidence interval (23), respectively.

Statistical Test

Using (23) (or (26)), we can perform a statistical test of the null hypothesis $H_0 : b_* = 0$ (or $r_* = 1$, resp.). The hypothesis H_0 means that X is irrelevant to the probability p of the outcome O, i.e., X is not a risk factor of O.

A statistical test of H_0 with significance level α is implemented by the following rule:

Accept H_0 if and only if $b_* = 0$ lies in the interval (23) (or equivalently $r_* = 1$ lies in the interval (26)).

This test is formulated in a slightly different manner. If $b_* = 0$, the definition (22) of u becomes

$$u = \frac{b}{\mathrm{SE}(b)}, \tag{27}$$

and u obeys the standard normal distribution. Then, the hypothesis H_0 is tested by the rule:

Accept H_0 if and only if $|u| < z(1-\alpha/2)$,

where u is defined by (27). Here, we can also use the statistic

$$u^2 = \frac{b^2}{\mathrm{SE}(b)^2}, \tag{28}$$

which obeys a chi-squared distribution with one degree of freedom. This is called the **Wald test**.

4. Multivariate Logistic Regression

The principle of logistic regression is generalized to multivariate situations, where simultaneous effects of exposures on an outcome are assessed. We assume that probability p of an outcome O may depend on exposure variables X_1, X_2, ⋯, X_r in the following form:

$$p = \frac{e^A}{1+e^A}, \tag{29}$$

where

$$A = a_* + b_{1*}X_1 + b_{2*}X_2 + \cdots + b_{r*}X_r. \tag{30}$$

The exponent A is a generalization of $a_* + b_*X$ in (21). This model is called a **multivariate logistic regression model**, and A is referred to as a *logit*. Population parameters a_*, b_{1*}, b_{2*}, ⋯, b_{r*} are estimated by a sample as in **Table 2** according to the method of maximum likelihood. If $b_{i*} \neq 0$ is shown for some i, the exposure measured by X_i is considered to be a risk factor of the outcome O. When the effect of X_i on O is estimated, the other variables X_j, $j \neq i$, are fixed. This means that possible confounding by X_j, $j \neq i$, is solved.

Table 2 The structure of data for multivariate logistic regression. Values of exposure variables X_1, X_2, ⋯, X_r and of an outcome variable Y are measured for n subjects. The value of X_j for the ith subject is denoted by x_{ji}. If y_i=1 (or y_i=0), it means that the outcome is observed (or not observed, respectively) for the ith subject

Subject	X_1	X_2	⋯	Xr	Y
1	x_{11}	x_{21}	⋯	x_{r1}	y_1
2	x_{12}	x_{22}	⋯	x_{r2}	y_2
⋮	⋮	⋮	⋮	⋮	⋮
n	x_{1n}	x_{2n}	⋯	x_{rn}	y_n

Interval Estimation and Statistical Tests

Software packages as SAS and SPSS output the optimal regression coefficients a, b_1, b_2, ⋯, b_r and their standard errors SE(a), SE(b_1), SE(b_2), ⋯, SE(b_r). Then, by means of the statistic (22) with b and b_* replaced by b_i and b_{i*}, respectively, we can make interval estimation of b_{i*} in the same way as (23), and a Wald test for a hypothesis $b_{i*} = 0$ is done. However, simultaneous Wald tests for two or more parameters may cause a problem in the interpretation of a significance level. To this end, we need a **likelihood ratio test** that we explain below.

The maximum likelihood procedure searches for optimal regression coefficients that maximize a quantity called a *likelihood*. We refer to the attained maximum as the *maximized likelihood value* L. The larger L is, the better the fit is. Then, we can compare goodness of fit for two regression models by means of L. Software packages usually output the value of L together with a Wald statistic.

Let us consider two logistic regression models

with logits

$$A_1 = a_* + b_{1*}X_1 + b_{2*}X_2 + \cdots + b_{r*}X_r, \quad (31)$$

$$A_2 = a_* + b_{1*}X_1 + b_{2*}X_2 + \cdots + b_{r+s*}X_{r+s}, \quad (32)$$

respectively, where the latter includes extra terms $b_{j*}X_j$, $j = r + 1, r + 2, \cdots, r + s$. If we put $b_{j*} = 0$ for $j \geq r + 1$ in A_2, we have $A_1 = A_2$. In this sense, the first model is a *reduced model* of the second. These models have their maximized likelihood values L_1 and L_2. Because the maximum likelihood procedure in the second model scans wider range of parameters than in the first model, L_2 is necessarily larger than or equal to L_1. If we find $L_1 = L_2$ with a certain significance level, it means that the extra terms in the logit A_2 do not really work, hence the exposures measured by X_j, $j \geq r + 1$, are not risk factors.

The significance test based on L_1 and L_2 is carried out by looking at a difference of *log likelihood statistics* $-2 \ln L_1$ and $-2 \ln L_2$:

$$l = (-2\ln L_1) - (-2\ln L_2) = 2\ln \frac{L_2}{L_1}, \quad (33)$$

where ln stands for natural logarithm. Under the null hypothesis

$H_0 : b_{j*} = 0$ for $j = r + 1, r + 2, \cdots, r + s$,

the statistic l approximately obeys a chi-squared distribution with s degrees of freedom. Then, the test of H_0 with significance level α is done according to the rule:

Reject H_0 if and only if $l > \chi^2_s(1 - \alpha)$,

where $\chi^2_s(1 - \alpha)$ denotes the $(1 - \alpha) \times 100$ percentile of a chi-squared distribution with s degrees of freedom.

By means of a likelihood ratio test, we can distinguish significant and nonsignificant variables. Therefore, it might seem possible to begin with a model that involves *all* the exposure variables and to eliminate nonsignificant variables as a result of the test. The maximization process for such a large model, however, is likely to become unstable, and computer outputs may be numerically unreliable. For this reason, it is recommended to remove variables at the outset that are clearly nonsignificant. To this end, we divide subjects into two groups with and without the outcome and compare the exposure variables of the groups. For example, if distributions of age for two groups are significantly different, the age variable should be involved in the model. On the other hand, if distributions of sex are not significantly different, the sex variable can be eliminated from the model. Note that these preliminary tests cannot be alternative to the logistic regression itself because of the problem of multiple comparisons. Note also that the preliminary tests must be conservative, that is, we have to make the significance level slightly larger, e.g., 10% so that potential risk factors may not escape our analysis.

Remark: If the logit A_2 contains only one extra term (i.e., $s = 1$), a Wald test is also available. In fact, for a large sample, a Wald test and a likelihood ratio test give almost the same result. For a small sample (the sample size < 20), however, a likelihood ratio test is recommended.

References

1. Watanabe H: Applications of Statistics to Medical Science, I. J Nippon Med Sch 2011; 78: 274–279.
2. Watanabe H: Applications of Statistics to Medical Science, II. J Nippon Med Sch 2012; 79: 31–36.
3. Rosner B: Fundamentals of Biostatistics, 7th ed., 2011; Brooks/Cole, Boston.
4. Armitage P, Berry G, Matthews JNR: Statistical Methods in Medical Research, 4th ed., 2002; Blackwell Science, Massachusetts.
5. Kleinbaum DG, Klein M: Logistic Regression, 3rd ed., 2010; Springer, New York.

(Received, January 6, 2012)
(Accepted, February 10, 2012)