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\textsuperscript{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\textsuperscript{14} are rich sources of information on medical statistics. For details of logistic regression, see\textsuperscript{1}

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 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be obtained data, that is, $x_i$ and $y_i$ denote the values of $X$ and $Y$, respectively, for the $i$th 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

$$f(X) = a + bX$$

is referred to as linear regression.

Regression Coefficients

Given experimental data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, 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

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

$$b = r_{xy} \frac{s_y}{s_x}. \quad (3)$$

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/)

J Nippon Med Sch 2012, 79 (2)
\[ r_{xy} = \frac{c_{xy}}{\text{SS}_x^{1/2}} \]  
where \( c_{xy} \) stands for a covariance

\[ c_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}), \]  

**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 \( \varepsilon \), associated with the \( i \)th subject and assume the relation between \( x \) and \( y \) as follows:

\[ y = a_* + b_* x + \varepsilon, \quad i = 1, 2, \ldots, n, \]  
where \( a_* \) and \( b_* \) are unknown population parameters. This means that the value \( y \) of \( Y \) has a deterministic part \( a_* + b_* x \) and a stochastic part \( \varepsilon \). Here, \( \varepsilon, \varepsilon_2, \ldots, \varepsilon_i \) are assumed to be independent random variables that obey the same normal distribution with mean 0 and variance \( \sigma^2 \). These assumptions are equivalent with the statement that \( y, y_2, \ldots, y_n \) obey independent normal distributions with mean \( a_* + b_* x \) and variance \( \sigma^2 \) for \( i = 1, 2, \ldots, n \), respectively. The sample values \( x_1, x_2, \ldots, x_n \) 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 \( \sigma^2 \) from a sample. The estimators for \( a_* \) and \( b_* \) are given by (2) and (3), and \( \sigma^2 \) is estimated by \( \text{SS}_{\text{residual}}/(n - 2) \), where \( \text{SS}_{\text{residual}} \), a residual sum of squares, is defined by

\[ \text{SS}_{\text{residual}} = \sum_{i=1}^{n} (a + bx_i - \bar{y})^2. \]  

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

\[ \text{SE}(a)^2 = \frac{\text{SS}_{\text{residual}}}{(n-2)n} \left(1 + \frac{\bar{x}^2}{s_x^2}\right), \]  
\[ \text{SE}(b)^2 = \frac{\text{SS}_{\text{residual}}}{(n-2)ns_x^2}, \]

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

\[ t = \frac{b - b_*}{\text{SE}(b)} \]  
obeys a \( t \) distribution with \( n - 2 \) degrees of freedom, hence the confidence interval of \( b \) with confidence level \( 1 - \alpha \) is written as

\[ b - t_{n-2}(1-\alpha/2)\text{SE}(b) < b < b + t_{n-2}(1-\alpha/2)\text{SE}(b), \]  

where \( t_{n-2}(1-\alpha/2) \) stands for the \( (1 - \alpha/2) \times 100 \text{th} \) 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 \( \text{SS}_{\text{regression}} \) by

\[ \text{SS}_{\text{regression}} = \sum_{i=1}^{n} (a + bx_i - \bar{y})^2, \]  

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

\[ F = \frac{\text{SS}_{\text{regression}}}{n - 2} \frac{1}{\text{SS}_{\text{residual}}} \]  

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

Reject \( H_0 \) if only if \( F > F_{\alpha} \) \( \cdot \) \( \cdot \) \( (\cdot) \),

where \( F_{\alpha} \cdot \) \( \cdot \) \( (\cdot) \) denotes the \( \alpha \times 100 \text{th} \) 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 \( F^2 = F \) 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

\[ f(X_1, X_2, \ldots, X_s) = a + b_1X_1 + b_2X_2 + \cdots + b_sX_s. \]  

The optimal coefficients \( a, b_1, b_2, \ldots, b_s \) 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 \( b_x \). 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

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

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 "\( a_x = 0 \) and \( b_x = 1 \)" 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, \( \kappa \) coefficient should be used.

3. Logistic Regression

We sometimes encounter problems to which the method of linear regression by the function (I) 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 (I) is inappropiate, 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.

**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 \( i \)th subject. If \( y_i = 1 \) (or \( = 0 \)), the probability \( f(x) \) of \( O \) for the \( i \)th subject should be estimated to be large (or small), i.e., near to 1 (or 0, resp.). In short, we expect that \( f(x) \) may be near to \( y_i \), i.e.,

\[
y_i \equiv f(x_i), \quad i = 1, 2, \ldots, 10. \tag{17}
\]

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

\[
f(X) = \frac{e^{\alpha X}}{1 + e^{\alpha X}}. \tag{18}
\]
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

\[
u = \frac{b - b^*}{SE(b)} \tag{22}
\]

approximately obeys the standard normal distribution, a confidence interval for \( b^* \) with confidence level \( 1 - \alpha \) is

\[
b - z(1 - \alpha/2)SE(b) < b^* < b + z(1 - \alpha/2)SE(b). \tag{23}
\]

where \( z(1 - \alpha/2) \) denotes the \( (1 - \alpha/2) \times 100\text{th} \) 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

\[
r^* = \frac{e^{a^*+b^*}}{e^{a^*}} = e^{b^*}. \tag{25}
\]

In view of the above formula for the odds ratio \( r^* \),
we see that $r^*$ has a point estimator $c^*$ and an interval estimate has the form

$$c^* < r^* < c^* + \alpha,$$  \hspace{1cm} (26)

where $b$ is the point estimator of $r_*$ and $b_*$ stand for the limits $b \pm z(1 - \alpha/2)SE(b)$ 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 $\alpha$ is implemented by the following rule:

Accept $H_0$ if and only if $b_* \neq 0$ 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

$$u = \frac{b}{SE(b)},$$  \hspace{1cm} (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

$$w^2 = \frac{b^2}{SE(b)^2},$$  \hspace{1cm} (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 $X_1, X_2, \ldots, X_r$ in the following form:

$$p = \frac{e^A}{1 + e^A},$$  \hspace{1cm} (29)

where

$$A = a_1 + b_1 X_1 + b_2 X_2 + \cdot \cdot \cdot + b_r X_r.$$

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 $a_1, b_{1*}, b_2, \ldots, b_r$ are estimated by a sample as in Table 2 according to the method of maximum likelihood. If $b_* \neq 0$ 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 $X_i, j \neq i$, is solved.

### Interval Estimation and Statistical Tests

Software packages as SAS and SPSS output the optimal regression coefficients $a, b_1, b_2, \ldots, b_r$ and their standard errors $SE(a), SE(b_1), SE(b_2), \ldots, SE(b_r)$. Then, by means of the statistic (22) with $b$ and $b_r$ replaced by $b_1$ and $b_*$, respectively, we can make interval estimation of $b_*$ in the same way as (23), and a Wald test for a hypothesis $b_* = 0$ 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

<table>
<thead>
<tr>
<th>Subject</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>\ldots</th>
<th>$X_r$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$x_{11}$</td>
<td>$x_{21}$</td>
<td>\ldots</td>
<td>$x_{r1}$</td>
<td>$y_1$</td>
</tr>
<tr>
<td>2</td>
<td>$x_{12}$</td>
<td>$x_{22}$</td>
<td>\ldots</td>
<td>$x_{r2}$</td>
<td>$y_2$</td>
</tr>
<tr>
<td>\vdots</td>
<td>\vdots</td>
<td>\vdots</td>
<td>\ddots</td>
<td>\vdots</td>
<td>\vdots</td>
</tr>
<tr>
<td>$n$</td>
<td>$x_{1n}$</td>
<td>$x_{2n}$</td>
<td>\ldots</td>
<td>$x_{rn}$</td>
<td>$y_n$</td>
</tr>
</tbody>
</table>
with logits

\[ A_1 = a_1 + b_1 X_1 + b_2 X_2 + \cdots + b_s X_s \]  

(31)

\[ A_s = a_s + b_1 X_1 + b_2 X_2 + \cdots + b_r X_r \]  

(32)

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

The significance test based on \( L_1 \) and \( L_s \) is carried out by looking at a difference of log likelihood statistics \(-2 \ln L_1\) and \(-2 \ln L_s\)

\[ l = (-2 \ln L_1) - (-2 \ln L_s) = 2 \ln \frac{L_s}{L_1}. \]  

(33)

where \( \ln \) stands for natural logarithm. Under the null hypothesis

\[ H_0 : b_s = 0 \quad \text{for} \quad j = r + 1, r + 2, \cdots, r + s, \]

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

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

where \( \chi^2(1 - \alpha) \) 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 \( A_1 \) 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


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