POLYTECHNIC UNIVERSITY
Department of Computer Science / Finance and Risk Engineering

Regression

K. Ming Leung

Abstract: Regression is a powerful method for making prediction of the value of a continuous response variable based on the values of one or more continuous predictor (or independent) variables.

Directory

Last Revision Date: December 26, 2007

Table of Contents

1. Introduction

Numeric prediction is the task of predicting continuous values of a dependent (or response) variable, Y, for given continuous values of one or more independent (or predictor) variables, $X_1, X_2, \ldots, X_n$. The response variable is what we want to predict for a given sample. The predictor variables are the attributes of interest describing the sample. The most widely used approach for numeric prediction is regression.

Several software packages exist to solve regression problems. Examples include SAS (www.sas..com), SPSS (www.spss.com), and S-Plus (www.insightful.com).

2. Linear Regression

In regression, we are looking for a relationship between the response variable, Y, and the predictor variables, $X_1, X_2, \ldots, X_n$. Such a relation is called a regression equation, which typically involves a number of unknown continuous parameters. The most widely used equation has the following linear form:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

where $\alpha, \beta_1, \ldots, \beta_n$ are $n+1$ unknown parameters of the linear model. These unknown parameters, known as regression coefficients, are to be determined from a set of training data samples.

We assume there are m samples in the training set. Each sample has n independent predictor values, $(x_1, x_2, \ldots, x_n)$, and a corresponding dependent response value, y. Applying the above equation to each of the given samples yields a set of linear equations:

$$y_j = \alpha + \beta_1 x_{1j} + \beta_2 x_{2j} + \ldots + \beta_n x_{nj} + \epsilon_j, \quad j = 1, \ldots, m.$$

where ϵ_j is the error of regression for sample j. This error term accounts for deviation from a relationship between the response variable, Y, and the predictor variables, $X_1, X_2, \ldots, X_n$ because of inaccuracies in the model and the random nature of real data.

Of course typically we have $m \gg n$. The case where there are more than one predictor variable (*i.e.* $n > 1$) is known as multiple linear regression. If there is only one predictor variable (*i.e.* $n = 1$), then we have the simpler case of straight-line linear regression.

2.1. Straight-Line Linear Regression

If we have only a single predictor variable, X, then we can drop one set of subscripts from our equations. The linear relation between X and Y is

$$Y = \alpha + \beta X.$$

The set of training data samples yield the set of linear equations:

$$y_j = \alpha + \beta x_j + \epsilon_j, \quad j = 1, \ldots, m.$$

The sum of the squares of the errors (SSE) for the entire data set is

$$\text{SSE} = \sum_{j=1}^{m} \epsilon_j^2 = \sum_{j=1}^{m} (y_j - \alpha - \beta x_j)^2.$$

The two regression coefficients can be found using the method of least square, which minimizes the SSE. Differentiating SSE with respect to α and β separately and setting the resulting equations to zero yield

two equations

$$\frac{\partial(\text{SSE})}{\partial\alpha} = -2\sum_{j=1}^{m}(y_j - \alpha - \beta x_j) = 0$$

$$\frac{\partial(\text{SSE})}{\partial\beta} = -2\sum_{j=1}^{m}(y_j - \alpha - \beta x_j)x_j = 0.$$

These two equations are linear in the unknowns α and β, and can be solved to obtain the solution

$$\alpha = \bar{y} - \beta\bar{x}$$

$$\beta = S_{xy}/S_{xx},$$

where $\bar{x}$ and $\bar{y}$ are the arithmetic mean values for variables X and Y in the training data set,

$$S_{xy} = \sum_{j=1}^{m}(x_j - \bar{x})(y_j - \bar{y})$$

$$S_{xx} = \sum_{j=1}^{m} (x_j - \bar{x})^2.$$

Also of interest is the quantity

$$S_{yy} = \sum_{j=1}^{m} (y_j - \bar{y})^2.$$

It is important to know the extend to which two variables, such as x an y, are correlated. The strength and direction of the relationship can be measured by a **correlation coefficient**, r, which is defined as

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \beta \sqrt{\frac{S_{xx}}{S_{yy}}}.$$

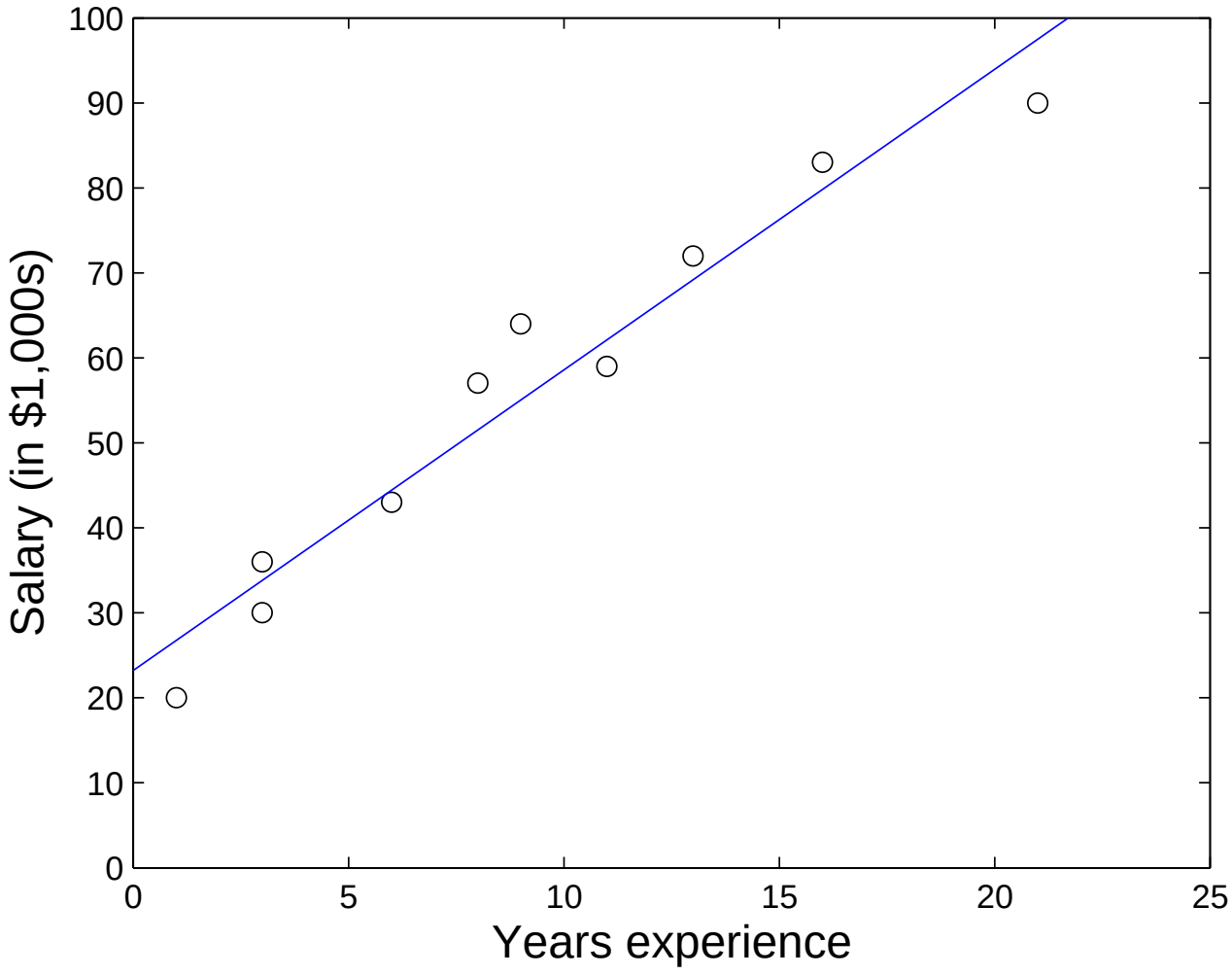
The correlation is defined only if S_{xx} and S_{yy} (which are basically the standard deviations) are both finite and nonzero. It is a corollary of the Cauchy-Schwarz inequality that the correlation cannot exceed 1 in absolute value. The correlation is 1 in the case of an increasing

linear relationship, -1 in the case of a decreasing linear relationship, and some value in between in all other cases, indicating the degree of linear dependence between the variables. The closer the coefficient is to either -1 or 1, the stronger the correlation between the variables. If the variables are independent then the correlation is 0, but the converse is not true because the correlation coefficient detects only linear dependencies between two variables.

- **Example: Straight-line Linear Regression**

Years experience (x)	Salary (y), in \$1000s
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

Suppose we are given the follow table of data where x is the number of years of work experience of a college graduate and y is the corresponding salary of the graduate. The data is 2-dimensional and can be easily graphed. The plot suggest a linear relationship between the two variables X and Y.


Salary (in $1,000s)
Years experience
0
10
20
30
40
50
60
70
80
90
100
0
5
10
15
20
25

Using straight-line regression, we find that $\beta = 3.5375$ and $\alpha = 23.2090$. Thus the equation of the least squares line is estimated by

$$y = 23.6 + 3.5x.$$

The correlation coefficient is found to be 0.9721, which indicates a very strong positive correlation between the salary of a college graduate and the number of years of work experience.

Using this above equation, we can now make prediction about the salary of any college graduate based on the number of years of work experience. For example, we find that the salary of a college graduate with, say, 10 years of experience is $\$58,583.70$.

2.2. Multiple Linear Regression

Multiple linear regression is an extension of straight-line linear regression so as to involve more than one predictor variable. It allows response variable y to be modeled as a linear function of n predictor variables or attributes, $A_1, A_2, \ldots, A_n$, describing a sample, $\mathbf{X} = (x_1, x_2, \ldots, x_n)$. Our training data set, T, contains m data points of the form $(\mathbf{X}_1, y_1), (\mathbf{X}_2, y_2), \ldots, (\mathbf{X}_m, y_m)$, where the $\mathbf{X}_i$ are

the n-dimensional training samples with associated continuous class labels, y_i.

Multiple linear regression problems are commonly solved using the general method of least squares with the use of statistical software packages, such as SAS, SPSS, and S-Plus, or more general numerical computing software such as Matlab.

3. Nonlinear Regression

How can we model data that does not show a linear dependence of the response variable to the predictor variables? One can generalize the above to include modeling by polynomial functions. By applying transformations to the variables, we can convert the nonlinear model into a multiple linear regression problem that can then be solved by the method of least squares.

For example, a cubic relationship given by

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

can be converted to linear form if we introduce independent variables

$$x_1 = x \quad x_2 = x^2 \quad x_3 = x^3$$

to obtain the linear multiple-regression model

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3.$$

By suitably transforming the independent variable, or transforming the dependent variable, or both, we can reduce a nonlinear relation to a linear one where linear regression can then be used. Some useful transformations can be found in this table.

Function form	**Transformations**	**Regression form**
Exponential:		
$Y = \alpha e^{\beta X}$	$Y^* = \ln Y$	Regress Y^* against X
Power:		
$Y = \alpha X^{\beta}$	$Y^* = \ln Y$, $X^* = \ln X$	Regress Y^* against X^*
Reciprocal:		
$Y = \alpha + \beta/X$	$X^* = 1/X$	Regress Y against X^*
Hyperbolic:		
$Y = X/(\alpha + \beta X)$	$Y^* = 1/Y$, $X^* = 1/X$	Regress Y^* against X^*

References

[1] M. Kantardzic, *Data Mining - Concepts, Models, Methods, and Algorithms*, IEEE Press, Wiley-Interscience, 2003, ISBN 0-471-22852-4. This is our adopted textbook, from which this set of lecture notes are derived primarily.