The theory and practice of linear regression

Hongyao Deng† & Xiuli Song‡

Yangtze Normal University, Chongqing, People's Republic of China†
Chongqing University of Posts and Telecommunications, Chongqing, People's Republic of China‡

ABSTRACT: A valid and efficient method to teach undergraduate courses in linear regression analysis is presented in this article. The method includes theory and practice parts, where interactive learning methodologies are created. It adopts case-study teaching, since this teaching method effectively integrates theoretical teaching and practical teaching. The lectures should be not an exhaustive review of regression methodology, but they should focus on how the regression models derived. Moreover, the teacher should pay more attention to the theoretical aspects of models rather than to their implementation using software. Students work in teams of three or four on a problem presented by teachers and choose relevant software to carry out their own projects. Feedback from students indicates that this method of teaching improves students' class attendance and greatly increases their interest in learning.

INTRODUCTION

Reality requires one to deal with a range of variables. Students can encounter many variables for a problem, including relationships between the variables. Commonly, these problems can be divided into two kinds according to their relationship. One is the problem of determining, which relationship can be expressed using the functions available. Another is not completely determined by using functions, but with random variables. This kind of relationship is called co-relationship. Regression analysis is a method to study the relationship between variables. So one builds regression models to help understand and explain the relationships. Regression models can also be used to predict actual outcomes.

The earliest form of regression was the method of least squares, which was published by Legendre in 1805 [1] and by Gauss in 1809 [2]. Legendre and Gauss both applied the method to problems of determining. The term *regression* was coined by Francis Galton in the 19th Century to describe a biological phenomenon [3][4]. However, Francis Galton applied the method to the problem of random variability. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average, and this is, known as regression toward the mean.

Today, regression methods continue to be an area of active research, and new methods have been developed for robust regression, regression involving correlated responses, such as time series and growth curves, regression in which the predictor or response variables are curves, graphs, images or other complex data objects, etc. There are many examples. Feilong and Yubo made an error analysis for the linear programming support vector regression problem in learning theory [5]. Jose M. proposed a machine learning algorithm for regression problems [6]. Qing, H. et al presented parallel extreme learning machine for regression based on MapReduce [7].

Chen Y. proposed assessing mathematics learning achievement using hybrid rough set classifiers and multiple regression analysis [8]. González-Recio and Forni proposed Genome-wide prediction of discrete traits using Bayesian regressions and machine learning [9]. Tsuruoka et al used logistic regression to study learning string similarity measures [10]. Therefore, almost all institutions in which relevant majors have been set offer a course in regression. For example, MIT offers open courseware [11]. In addition, a large number of open on-line teaching resources can also be found, such as applied regression analysis at Columbia Business School for the MBA major.

In 2011, a statistics major was established in the Mathematics and Computer School at Yangtze Normal University, Chongqing, China. In this year, the School first offered the course for students whose major came from computer science or mathematics. It is compulsory for mathematics students, while it is an elective for computer science students.

Its objective is to strengthen their statistical skills. One of the authors received and accepted the teaching load. Hence, sharing the teaching experience is the objective of this article.

THEORY TEACHING

The theory-oriented lectures cover single linear regression and multiple linear regressions. Students learn what regression is, how to create its mode, how to estimate the parameters of the model (Estimation Using Least Squares), understanding the assumptions of establishing the conditions for the model, what the regression coefficients are, how to compare the models, and predicting and controlling using regression model. Teachers start theirs lectures with a discussion of simple regression, then, move on to multiple linear regression. This is quite reasonable from a pedagogical point of view, since simple regression has the great advantage of being easy to understand graphically.

Students should place a lot of emphasis on the simple linear regression analysis and understanding its mathematical expressions and be open to more sophisticated concepts. It is difficult for students to study multiple linear regression analysis. However, it is a primary tool in the analysis of real data. Thereby, single linear regression is taught in six sessions, while multiple linear regression requires four sessions. A session is 90 minutes' duration.

Single Linear Regression Model

Single linear model describes a linear relationship between two variables. One is called the target, response or dependent variable, and is usually represented by y. Another is called the predicting or independent variables, and is usually represented by x. Given $(x_1, x_2, \cdots, x_N)$, the simple linear regression model is described as:

$$\begin{cases} y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \\ \varepsilon_i \sim N(0, \sigma^2) \text{ and } \operatorname{cov}(\varepsilon_i, \varepsilon_j) = 0 \text{ when } i \neq j \end{cases} \tag{1}$$

where the data, $\{x_i, y_i\}$, represent a random sample from a larger population, which consist of n set of observations, the β coefficients are unknown parameters, and ε_i are random error or disturbance terms.

Least Square Estimation

A primary goal of a regression analysis is to estimate the relationship between the predictor and the target variables or equivalently, to estimate the unknown parameter β. This requires a data-based rule or criterion that will give a reasonable estimate. The standard approach is least squares regression which is a convex optimisation problem with no constraints. The objective is a sum of squares of terms of the form that are chosen to minimise:

$$\sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 \tag{2}$$

The scatter diagram gives a graphical representation of least squares, which can help students to understand regression graphically. If the fitted regression equation has been obtained, it is a line given by:

$$\hat{E}(y) = \hat{\beta}_0 + \hat{\beta}_1 x \tag{3}$$

Residuals are defined as the difference between the observed value y_i and the fitted value $\hat{y}_i$. Equation (2) minimises the sum of squares of the residuals if the coefficients β take as the fitted coefficient $\hat{\beta}$. By minimising Equation (2), the regression coefficients are obtained by:

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = S_{XY} / S_{XX} \end{cases} \tag{4}$$

where $\bar{x} = \frac{1}{n}\sum x_i$, $\bar{y} = \frac{1}{n}\sum y_i$, $S_{XX} = \sum (x_i - \bar{x})^2$, $S_{XY} = \sum (x_i - \bar{x})(y_i - \bar{y})$.

Hypothesis Tests for the Simple Regression Model

Hypothesis testing is used to carry out inferences about the regression parameters β_1 using data from a sample set. The objective is to test the overall significance of the regression. It is necessary for the lectures to place the emphasis on

three main steps. Firstly, formulate a null hypothesis, $H_0 : \beta_1 = 0$, and an alternative hypothesis, $H_1 : \beta_1 \neq 0$. Secondly, build a statistic to test the hypothesis made. Thirdly, define a decision rule to reject or not to reject the null hypothesis. In addition, a simple specific example should be given in order to explain the concept of hypothesis test theory. Moreover, the role of the coefficient β_1 must be explained to students.

It can be proved from Equation (4) that the estimated expected change in the target variable is associated with the predicting variable. Therefore, it is significant for the regression if the null hypothesis H_0 is rejected. Otherwise, it shows that there is no linear association. This is that the regression equation is not significant. So, a few of cases can arise:

- The predictor variable *x* has no significant effect on the target variable *y*. In this case, the variable *x* should be rejected.
- The predictor variable *x* has a significant effect on the target variable *y*. The relation between them is not linear but non-linear.
- In addition to predictor variable *x*, there are other predictor variables to significantly influence target variable *y*, thereby, the effect on the variable *y* is impaired. In this case, multiple regression models must be considered.

Building a statistic to test the hypothesis takes into account the H_0 and the sample data. In practice, as σ^2 (variance of ε) is always unknown, the distributions t and F will be used. The $F-test$ has the form:

$$F = \frac{S_{YY}}{SS_{reg}/(n-2)} \sim F(1, n-2) \tag{5}$$

where $S_{YY} = \sum (y_i - \bar{y})^2$ termed the corrected total sum of square, which measures the variation of the y_i values around their mean y, $SS_{reg} = \sum (\hat{y} - \bar{y})^2$ is the residual sum of square, which explained variation attributable to the relationship between x and y.

The *t*-test is:

$$t = \frac{S_{XX}}{\sqrt{SS_{reg}/(n-2)}} \sim t(n-2) \tag{6}$$

Additionally, the sample relationship coefficient can also test the hypothesis. Its formula is:

$$r = \frac{SS_{reg}}{S_{YY}} = \left(\frac{S_{XY}}{S_{XX}S_{YY}}\right)^2 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = R \tag{7}$$

Note that R is the relationship coefficient in statistics. The three methods are not different in nature.

Decision rules are constructed making the probability of H_0 small. Common values for significance level are 0.10, 0.05 and 0.01, although sometimes 0.001 is also used. Of course, a specific example should be given to help students learn how to compute the probability.

As for multiple linear regression, it is conceptually similar to the simple linear regression. The distinguishing feature between multiple linear regression and single linear regression is the presence of more than one predictor variable in the multiple regression model. Therefore, teaching methods are similar to the single linear regression.

PRACTICE EXERCISE

Theory-oriented lectures on linear regression are difficult to teach and to be understood. It is a deliberate choice for teachers to create interactive learning methodologies. A major aspect of course is the opportunity to carry out a practical statistical analysis project of students' own making. There is some software to apply to the linear regression analysis, such as Excel, SPSS, MATLAB, R, etc.

Students choose appropriate software to carry out their projects according to their majors. The practice may be divided into two stages. One is to work in teams; students work in teams of three or four on a problem presented by the professor.

The goal of the project is to develop a useful statistical model for a specific real problem. The teams submit a test report and give an oral presentation of their results. Another objective is to work individually. In this stage, the question presented is comprehensive. The goal is not only to develop a useful statistical model, but also to apply this model to reality. Hence, students exercise their spirit of cooperation but, on the other hand, cultivate their ability to accomplish a task.

The combination of theory and practice improve students' application and theoretical knowledge. Here, an MS Excel example to implement linear regression is presented. The contents of linear regression in MS Excel cover creating an initial scatter plot, creating a linear regression line (trend line), using the regression equation to calculate the regression coefficients (slope and intercept) and using the hypothesis tests to estimate fit. The following introduce some methods to compute linear regression.

Using MS Excel Function

Microsoft Excel has built in functions SLOPE and INTERCEPT to calculate linear regression slope and intercept. The syntax is SLOPE(known_y's, known_x's) and INTERPRET(known_y's, known_x's), where the known_y's are dependent and known_x's are independent. The function CORREL can calculate the sample relationship coefficient.

Except the functions above, there is a function LINEST to perform linear regression analysis. Its syntax is LINEST(known_y's, known_x's, const, stats) where const is a logical value specifying whether to force the constant b to equal 0, if const is TRUE or omitted, b is calculated normally, otherwise, b is set equal to 0, stats is also a logical value specifying whether to return additional regression statistics, If stats is TRUE, LINEST returns the additional regression statistics, If stats is FALSE or omitted, LINEST returns only the interpret-coefficients β_1 and the constant β_0.

Using MS Excel Data Analysis Add-Ins

Microsoft Excel has excellent Data Analysis menu and one of them contain regression. The approach is more simple than using functions. MS Excel has automated the procedure to compute regression by drop-down menus and dialogue boxes. To use the data analysis menu: Open your MS Excel, in the Microsoft Office button, click Excel Options item and pop up the Excel Options dialog, click Add-Ins, click GoTo button and pop up Add Macro-command dialog, check the Analysis ToolPak in the Add-ins available box and, then, press OK button. Next time when you open the Tools menu again, you will see Data Analysis in the bottom of Tools menu.

Other Some Ways

Using MS Excel graph trend can solve regression problem. It is easy to learn how to use MS Excel graph trend. After plotting the data in the XY type graph, the regression line can be also obtained directly from the graph. The R-squared value can also be displayed on the chart. Generalising inverse matrix is another way to solve regression problem, even manual calculation.

This practice is very useful. It may give students a better preparation for further numerical analysis of curvilinear regression, multiple linear regressions and non-linear regression. It requires students to use matrix formula. MS Excel has build in function MMULT, MINVERSE and TRANSPOSE. The function MMULT returns the results of multiplication of two matrices, MINVERSE returns the inverse matrix and Transpose returns the transposed matrix. Functions are terminated by CTRL+SHIFT+ENTER.

CASE-STUDY TEACHING

Case-study teaching is a kind of teaching method that effectively integrates theoretical teaching and practical teaching. There is a specific case in teaching. The objective of the course is linear regression analysis. The software of application is MS Excel 2007. The task is to find a linear equation that describes a data set and measures the strength of the regression relationship. These tasks using MS Excel Data Analysis Add-Ins finish the case.

Suppose that the data have been entered into the spreadsheet, and consist of two columns of numbers that will be representing the Y and the X Range. The Y represents the dependent variable while the X is the independent variable. Firstly, open the Regression Analysis tool. Define the Y and X Range: in the Regression Analysis box, click inside the Y Range box, then, click and drag the cursor in the Y Range field to select all the numbers analysed, repeat the previous step for the X Range.

Modify the settings if desired: Choose whether or not to display labels, residuals, residual plots, etc, by checking the desired boxes, here one only checks the Display Labels. Designate where the output will appear: one can either select a

particular output range or send the data to a new workbook or worksheet, here one selects a particular output range. Finally, click OK, and the summary of our regression output will appear where designated, as can be seen in Figure 1:

Data	
X	Y
0.25	2.57
0.37	2.31
0.44	2.12
0.55	1.92
0.60	1.75
0.62	1.71
0.68	1.60
0.70	1.51
0.73	1.50
0.75	1.41
0.82	1.33
0.84	1.31
0.87	1.25
0.88	1.20
0.90	1.19
0.95	1.15
1.00	1.00

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.9937442							
R Square	0.9875275							
Adjusted R Square	0.9866366							
Standard Error	0.0424592							
Observations	16							
ANOVA								
	df	SS	MS	F	Significance F			
Regression	1	1.998336007	1.998336	1108.4715	9.88962E-15			
Residual	14	0.025238993	0.0018028					
Total	15	2.023575						
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	lower 95.0%	upper 95.0%
Intercept	2.9909158	0.045546789	65.666886	7.80E-19	2.893227651	3.0886039	2.89322765	3.088603944
0.25	-2.016637	0.0605711	-33.293716	9.89E-15	-2.146549077	-1.8867249	-2.1465491	-1.886724899

Figure 1: The left two columns are the data consisting of observed variable X and dependent variable Y. The right solid box is the summary output for regression analysis.

FEEDBACK AND ASSESSMENT

In the spring semester of the 2012/2013 term, a total of 60 junior students came from two majors, computer science and mathematics, joined the course at the Yangtze Normal University, Chongqing, China. These students already had joined some previous courses, such as mathematical analysis, advanced algebra, probability and statistics, etc. They also demonstrated a certain level of operating computer software capability. A survey questionnaire with six feedback questions was administered to all of these students (shown in Table 1).

Table 1: Feedback survey questions.

Questions		Majors		Mean
		Computer science (students 30)	Mathematics (students 30)	
1	I became interested in data analysis through this class.	27 (90%)	28 (93%)	92%
2	I understand the concept of linear regression.	24 (80%)	26 (87%)	83%
3	I performed the linear regression using MATLAB program.	27 (90%)	26 (87%)	88%
4	I performed the linear regression using MS Excel.	29 (97%)	29 (97%)	97%
5	I think the practice is useful to understand the concept of linear regression.	30 (100%)	30 (100%)	100%
6	I can apply my knowledge of mathematics by performing the experiment.	24 (80%)	25 (83%)	82%

There are too many formulations about regression analysis. It is often difficult to understand these concepts of regression analysis, said some students. As can be seen from student feedback, only 80% of computer science students can understand regression, while 87% of mathematics students can do so. However, almost students finished their experiments, and all of them think the practice is very useful to understand those concepts. This information shows that students' interest will be enthused if theory is combined with practice, and as long as theory explanation is not ignored.

CONCLUSIONS

Regression analysis is not only a statistical process for estimating the relationships between variables, but it is also widely used for prediction and forecasting. It is difficult to acquire this knowledge, because it is necessary for undergraduates to have fundamental mathematics and operating computer skills. It includes many techniques for modelling and analysing several variables.

However, the lectures should be not an exhaustive review of regression methodology, but should focus on how the regression models are derived. In the teaching process, giving hands-on experience to students is necessary. The teacher should pay more attention to the theoretical aspects of models rather than to their implementation using software. In the process of teaching, these will help to improve the teaching effect.

REFERENCES

1. Legendre, A.M., Nouvelles méthodes pour la détermination des orbites des comètes. Paris: Firmin Didot, *Sur la Méthode des moindres quarrés* appears as an appendix (1805).
2. Gauss, C.F., Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientum. Ghent University (1809).
3. Mogull, R.G., *Second-Semester Applied Statistics*. Kendall/Hunt Publishing Company, 59-59 (2004).
4. Francis, G., Kinship and Correlation. *Statistical Science*, 4, **2**, 80-86 (1989).
5. Feilong, C. and Yubo, Y., Learning errors of linear programming support vector regression. *Applied Mathematical Modelling*, 35, 1820-1828 (2011).
6. José, M.M. and Pablo, E.M., Regularized extreme learning machine for regression problems. *Neurocomputing*, 74, 3716-3721 (2011).
7. Qing, H., Tianfeng, S., Fuzhen, Z. and Zhongzhi S., Parallel extreme learning machine for regression based on Map Reduce. *Neurocomputing*, 102, 52-58 (2013).
8. Chen, Y.S. and Cheng, C.H., Assessing mathematics learning achievement using hybrid rough set classifiers and multiple regression analysis. *Applied Soft Computing*, 13, 1183-1192 (2013).
9. González-Recio, O. and Forni, S., Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. *Genetics Selection Evolution*, 43, **7** (2011).
10. Tsuruoka, Y., McNaught, J., Tsujii, J. and Ananiadou, S., Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. Bioinformatics, 23, **20**, 2768-2774 (2007).
11. Massachusetts Institute Technology, Statistical Thinking and Data Analysis (2011), http://ocw.mit.edu/courses/sloan-school-of-management/15-075j-statistical-thinking-and-data-analysis-fall-2011/