USING MULTIPLE REGRESSION IN SOCIAL SCIENCES RESEARCH: SOME IMPORTANT ASPECTS

Lay Yoon Fah

ABSTRACT

This article discusses one of the most commonly used statistical methods in studying the relationship between dependent and independent variables in social sciences research. The purpose of using multiple regression and types of data suitable for multiple regression analysis are discussed. Some important aspects to be considered when multiple regression analysis is used will be discussed in detail. These aspects are variables selection method (i.e. forward selection, backward elimination and stepwise), multicollinearity, tolerance, variance inflation factor, influence statistics (DFFIT and DFBETA), Leverage, Cook's distance, standardized regression coefficient (β), coefficient of determination (R²), assumptions such as normality, linearity, homoscedasticity and independence. Figures and tables are illustrated to give better picture of the concepts described.

INTRODUCTION

Multiple Regression (MR) is a statistical method for studying the relationship between a single dependent variable and one or more independent variable(s). MR is used both for predicting outcomes (prediction) and for investigating the causes of outcomes (causal analysis). Dependent variable is a variable being predicted or explained by the set of independent variables whereas independent variable is variable selected as predictors and potential explanatory variables of the dependent variable (Hair, Anderson, Tatham & Black, 1998).

Some terms are used interchangeably to represent dependent variable and independent variable. For example, dependent variable is also known as response variable and outcome variable whereas independent variable is also called predictor variable, explanatory variable, regressor variable or covariates.

(a) Types of data

Multiple Regression can be used to analyze interval or ratio data. Ordinal variables are inappropriate for MR because the linear equation, to be meaningful, requires information on the magnitude of changes. In practice, however, ordinal variables are used quite often in regression analysis (Miles & Shevlin, 2001). If we use such variables, we are implicitly assuming that an increase (or a decrease) of one unit on
the scale means the same no matter where we start. This might be a reasonable approximation in many cases. On the other hand, MR allows us to statistically control for measured variables, but this control is never as good as a randomized experiment.

Dummy variables (0,1) are perfectly right as independent variables in a MR. Thus, the coefficients for dummy variables usually can be interpreted as differences in means on the dependent variable for the two categories of the independent variables, controlling for other variables in the regression model.

(b) Variable(s) Selection Method

(i) Forward selection

This variable selection method starts with a regression model that contains only the constant term. At each step, we add the variable that results in the largest increase in multiple $R^2$, provided that the change in $R^2$ is large enough for us to reject the null hypothesis that the true change is zero (0), using a preset significance level (the default criterion is an observed significance level of .05 or less). We stop entering variables into the model when there are no more variables that result in a significant increase in $R^2$.

(ii) Backward elimination

This variable selection method starts with a regression model that contains all of the independent variables. At each step, we remove the variables that changes $R^2$ least, provided that the change is small enough so that we can’t reject the null hypothesis that the change is zero (0), using a preset significance level (the default criterion is an observed significance level of .10 or larger). We stop removing variables when removal of any variable in the model results in a significant change in $R^2$.

(iii) Stepwise

This is the most commonly used method for model building. It is used to get a parsimonious model which can explain most of the variance in dependent variable by using the least number of independent variable. It is a combination of forward selection and backward elimination. It resembles forward selection except that after we enter a variable into the model, we remove any variables already in the model that are no longer significant predictors. This means that variables whose importance diminishes as additional predictors are added are removed.

We select the first two variables for entry the same way as in forward selection. Then we examine the variables in the model to see if either of them meet the removal criteria. If so, we remove it from the model. At each step, we enter a new variable using the same rules as in forward selection; then we examine the variables
Using Multiple Regression in Social Sciences Research

already in the model for removal, using the same rules as in backward elimination. The significance level for entering a variable (.05 or less) should be smaller than the significance level for removing a variable (.10 or larger).

(c) Multicollinearity

Extreme multicollinearity means that at least two of the independent variables in a regression equation are perfectly related by a linear function (e.g. \( y = a + b_1 x_1 + b_2 x_2 + u; x_1 = 3 + 2x_2 \)). Near extreme multicollinearity simply means that there are strong (but not perfect) linear relationships among the independent variables. If the regression model has only two independent variables, near extreme multicollinearity occurs if the two variables have a correlation that is close to +1 or -1. The closer to +1 or -1, the greater the associated problems.

Multicollinearity has nothing to do with the dependent variable, it is a characteristic of the relationships among the independent variables. Regress each independent variable on all the other independent variables and look for a high \( R^2 \) in any of these regressions. Suppose you have three independent variables, \( x_1, x_2 \) and \( x_r \). Pick one of them, say \( x_r \), and regress it on \( x_1 \) and \( x_2 \). If the \( R^2 \) for that regression is near +1 (\( R^2 > .60 \)) then \( x_r \) is said collinear with \( x_1 \) and \( x_2 \).

When two or more independent variables are highly collinear, it can appear that neither of them affects the dependent variable, but when either is excluded from the model, the remaining variable may have a highly significant effect. Multicollinearity is not so serious when the main goal of the regression analysis is to predict the dependent variable. Multicollinearity only affects the coefficient estimates for those variables that are collinear. It does make it more difficult to reliably estimate the coefficients of those variables that are collinear. Thus, it's impossible to get separate estimates for the collinear coefficients, \( b_1 \) and \( b_2 \). If two variables are perfectly correlated, when you hold one constant, the other must be constant as well. Hence, it's impossible to separate their effects on the dependent variable. Some of the recommended solutions for multicollinearity are: Delete one or more variables from the model; combine the collinear variables into an index; estimate a latent variable model; and perform joint hypothesis tests.

(i) Tolerance (T)

Tolerance is the extent to which an independent variable cannot be predicted by the other independent variables. Tolerance is calculated as \( 1-R^2 \) where the variable being considered is used as the dependent variable in a regression analysis and all other variables are used as independent variables. Tolerance varies between zero (0) and one (1). A tolerance value of zero (0) for a variable means that it is completely predictable from the other independent variables and that there is a perfect collinearity. If a variable has a tolerance value of one (1), this means that the variable is completely
uncorrelated with the other independent variables. Thus, we require a tolerance value close to one (1).

(ii) **Variance Inflation Factor (VIF)**

Variance Inflation Factor is closely related to the Tolerance. VIF is calculated by the formula \( VIF = \frac{1}{\text{Tolerance}} \). It relates to the amount that the standard error of the variable has been increased because of collinearity. The increase in standard error (SE) is equal to the square root of the VIF. Thus, we require a VIF value less than two (2).

(d) **Influence statistics**

Influence statistics is used for the identification of unusual observations and tells us how much the regression results would change if each individual observation was deleted from the analysis. Examine the data carefully for cases with large influence statistics to make sure there are no errors. If the deletion of an observation produces a big change, the observation is said to be influential.

The unstandardized influence statistics tells us how much the actual coefficient would change if a particular case was deleted. The standardized influence statistics divide the unstandardized influence statistics by the standard error of the coefficient. This makes it possible to compare the values across different variables.

(i) **DFFIT**

DFFIT tells us how much the predicted value for each observation would change if that observation were deleted from the regression analysis. It is the change in the predicted value of the dependent variable if the current case is omitted from the calculations. Examines the change in the predicted value of a case when that case is excluded i.e. the difference between the predicted value for a case when the case is included in the model, and the predicted value that would be calculated for that case when the case is removed from the model.

(ii) **DFBETA**

DFBETA tells us how much the coefficient for each of the independent variables would change if the observation were deleted. It is a new variable for each term in the regression model, including the constant, containing the change in the coefficient for that term if the current case were omitted from the calculations. It is the difference between the value of beta (i.e. the slope when the case is included, and the value of beta when the case is excluded). A rule of thumb is to look at cases with absolute values greater than \( \frac{2}{\sqrt{N}} \), where \( N \) is the number of cases. When we have a large number of cases, it's unlikely that removal of a single case will change the actual value of a coefficient very much.

76
(e) Leverage

Leverage is a measure of how greatly the current case influences the fit of the regression model. Leverage is calculated using only the values of independent variables. It is used to identify cases with unusual combinations of values of the independent variables. Leverage measures how far the values for a case are from the means of all of the independent variables. Leverage values range in value from zero (0) to close to one (1). Cases with high leverage values may have a large impact on the estimates of the regression coefficients. A rule of thumb is to look at cases with leverage values greater than $2p/N$, where $p$ is the number of independent variables in the model and $N$ is the number of cases.

(f) Cook's distance

Cook's distance is a measure of how much the residuals of all cases would change if the current case were omitted from the calculations. Cook's $D$ uses both the independent variable and dependent variable in its calculation. Cook's $D$ uses both the value of studentized residual and that of the leverage statistics to calculate a distance. It measures the change in all of the regression coefficients when a case is eliminated from the analysis. Cook's distances greater than one (1) usually deserve scrutiny.

(g) Standardized regression coefficient ($\beta$)

Standardized regression coefficient tells us how many standard deviation of the dependent variables changes for an increase of one standard deviation in a particular independent variable. All standardized coefficients are in the same metric, we can compare them across different variables. Standardized regression coefficient ($\beta$) and standard error are used to calculate the confidence interval where the true coefficient lies.

(h) Coefficient of determination ($R^2$)

Coefficient of determination or commonly known as $R^2$ is the most often used statistics to measure how well the dependent variable can be predicted from knowledge of the independent variables. Coefficient of determination for a regression is equal to the squared correlation between the dependent variable and the predicted value of the dependent variable, based on the estimated regression model.

The adjusted $R^2$ is a modification of the $R^2$ that adjusts for the number of independent variables. The adjusted $R^2$ is always less than or equal to the original $R^2$, and the discrepancy gets longer as the number of independent variables increases.

The basic idea behind $R^2$ is to compare two quantities i.e. The sum of squared errors (SSE) produced by the least squares equations that we are evaluating and the
sum of squared errors for a least squares equation with no independent variables (just the intercept). When an equation has no independent variables, the least squares estimate for the intercept is just the mean (average) of the dependent variable. That implies that our predicted value for every case is the mean. Thus, the formula to calculate coefficient of determination is $R^2 = 1 - \frac{\text{SSE (residual)}}{\text{SSE (mean only)}}$

(i) Violation of assumptions

(ii) Normality

For each value of the independent variable, the distribution of the values of the dependent variable must be normal. If the regression assumptions are met, the distribution of the ordinary residuals and the standardized residuals should be approximately normal. For samples larger than 30 cases or so, the distribution of studentized residuals should be normal as well.

If the data are a sample from a normal distribution, we expect the points to fall more or less on a straight line in the Q-Q plot of standardized residuals. For each of the points in the Q-Q plot, it shows the distance from the observed point to the line. On the other hand, if the data are a sample from a normal population, the points in the detrended normal plot should fall randomly in a band around zero (0). Statistical test of normality (Kolmogorov-Smirnov) which is more stringent can also be used to test normality. When the sample is small, the test is not very powerful, i.e. we often won’t reject the hypothesis of normality (i.e. the distribution is normal) even when it is incorrect. If the sample size is large, the test of normality may lead us to reject the normality assumption (i.e. the distribution is not normally distributed) based on small departures that won’t affect the regression analysis. As long as the normality assumption is not badly violated, the results of regression analysis will not be seriously affected.

(ii) Linearity

The relationship between the dependent and the independent variable must be linear in the population. The means of the distributions of the dependent variable must fall on a straight line. We can evaluate the linearity assumption by plotting the studentized residuals against the predicted values. If the relationship between the dependent variable and the independent variable is not linear, we will see a curve in the plot.

(iii) Homoscedasticity

The variance of the distribution of the dependent variable must be the same for all values of the independent variable. We can plot the studentized residuals against the predicted values to check for homoscedasticity. If the variance is constant, we
won't see any pattern in the data points i.e. the residuals appear to be randomly scattered around a horizontal line through zero (0).

(iv) Independence

All of the observations must be independent i.e. inclusion of one case in the sample must not influence the inclusion of another case or the value of one observation is in no way related to the value of another observation. We can check the independence assumption by plotting the studentized residuals against the sequence variable.

Durbin-Watson tests can also be used to see if adjacent observations are correlated. This statistics ranges in value from 0 to 4. If there is no correlation between successive residuals, the Durbin-Watson statistic should be close to 2. Values close to 0 indicate that successive residuals are positively correlated, while values close to 4 indicate strong negative correlation. As a rule of thumb, if the observed value is between 1.5 and 2.5, we need not worry.

Example

In a non-experimental quantitative research (Lay, 2006) to investigate the influence of science process skills, logical thinking abilities, attitude toward science and locus of control on science achievement among Form 4 students in the Interior Division of Sabah Malaysia, the dependent variable being predicted / explained is students' science achievement whereas independent variables i.e. variables selected as potential predictors or explanatory variables of students' science achievement are science process skills, logical thinking abilities, attitude toward science and locus of control.

Instruments used for collecting interval / ratio data to investigate the influence of science process skills, logical thinking abilities, attitude toward science and locus of control on students' science achievement are as follows:

i) Basic Science Process Skills Test (BSPST) - Interval/ratio data

ii) Integrated Science Process Skills Test (ISPST) - Interval/ratio data

iii) Group Assessment of Logical Thinking Abilities (GALT) - Interval/ratio data

iv) Attitude Toward Science In School Assessment (ATSSA) - Ordinal/Interval data

v) Intellectual Achievement Responsibility Questionnaire (IAR) - Interval/ratio data

vi) Science Achievement Test (SAT) - Interval/ratio data
Lay Yoon Fah

Table 1:
Multiple regression results for science process skills, logical thinking abilities, attitude toward science and locus of control on science achievement

<table>
<thead>
<tr>
<th>Predictor variables</th>
<th>B</th>
<th>SE</th>
<th>β</th>
<th>ΔR²</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>2.245</td>
<td>1.162</td>
<td>-</td>
<td>-</td>
<td>1.932</td>
<td>.054</td>
</tr>
<tr>
<td>Process</td>
<td>.340</td>
<td>.029</td>
<td>.534</td>
<td>.339</td>
<td>11.609*</td>
<td>&lt;.0005</td>
</tr>
<tr>
<td>Logic</td>
<td>.258</td>
<td>.115</td>
<td>.103</td>
<td>.008</td>
<td>2.235*</td>
<td>.026</td>
</tr>
<tr>
<td>Locus</td>
<td>-</td>
<td>-</td>
<td>.067</td>
<td>-</td>
<td>1.636</td>
<td>.103</td>
</tr>
<tr>
<td>Attitude</td>
<td>-</td>
<td>-</td>
<td>.036</td>
<td>-</td>
<td>.872</td>
<td>.384</td>
</tr>
</tbody>
</table>

*p < .05; Science = 2.245 + .340 Process + .258 Logic
Multiple R = .589
R² = .347
Adjusted R² = .344
SEE = 4.3163
F(2, 397) = 105.625; p < .0005

Results in Table 1 showed that science process skills and logical thinking abilities significantly contributed to students' science achievement (F(2, 397) = 105.625, p < .0005). Based on the R² value, we can conclude that these two predictor variables explained 34.7% of the variance in students' science achievement. In this matter, science process skills (R² = 33.9%, b = .534, t(400) = 11.609, p < .0005) contributed 33.9% of the variance in student's science achievement whereas logical thinking abilities only contributed 0.8% of the variance in students' science achievement (R² = 0.8%, b = .103, t(400) = 2.235, p = .026).

Figure 1 shows the standardized DFFIT of students' science achievement (dependent variable).
Using Multiple Regression in Social Sciences Research

Figure 1: Standardized DFFIT of science achievement

Figure 2 shows the standardized DFBETA of science process skills (independent variable) whereas Figure 3 shows the standardized DFBETA of logical thinking abilities (independent variable).
Figure 3: Standardized DFBETA of logical thinking abilities

Figure 4 shows the centred leverage value of each observation.

Figure 4: Leverage

Figure 5 shows the Cook's distance of each observation.
Using Multiple Regression in Social Sciences Research

Figure 5: Cook's distance

Figure 6, Figure 7, Figure 8 and Figure 9 shows the histogram of regression standardized residual of science achievement, P-P plot of regression standardized residual, stem-and-leaf plot of studentized deleted residual, and boxplot of studentized deleted residual respectively.

Figure 6: Histogram of regression standardized residual

Histagram
Dependent Variable: Pencapaian sains

Regression Standardized Residual

Std. Dev = .98
Mean = -.02
N = 445.00

Figure 6: Histogram of regression standardized residual
P-P Plot of Regression Standardized Residual

Dependent Variable: Pencapaian sains

Observed Cum Prob

Figure 7: P-P plot of regression standardized residual

Studentized Deleted Residual Stem-and-Leaf Plot

Frequency Stem & Leaf

3.00   -2.  56
5.00   -2.  06
20.00  -1.  555667776
53.00  -1.  000000011111112222233334444
61.00  -0.  5556666666677777888888999999
81.00  -0.  0000000001111111111222223333333334444
85.00  0.  0000000001111111111222222333333333444444
64.00  0.  55555555555666777777888888899999999
48.00  1.  000000001111111112222222333333333344444444
18.00  1.  555566786
5.00   2.  06
2.00 Extremes  (>=3.3)

Stem width:  1.00000
Each leaf:  2 case(s)
6 denotes fractional leaves.

Figure 8. Stem-and-leaf plot of studentized deleted residual
Figure 9: Boxplot of studentized deleted residual

Figure 10 shows that there exists a linear relationship between the science achievement (dependent variable) and science process skills (independent variable).

Partial Regression Plot

Dependent Variable: Pencapaian sains

Figure 10: Linearity
Figure 11 shows homoscedasticity where the residuals appeared to be randomly scattered around a horizontal line through zero.

Figure 12 shows there is no correlation between successive residuals since the Durbin-Watson statistic is 1.722 i.e. close to 2.

Durbin-Watson statistics: 1.722
CONCLUSION

This article has discussed some of the important aspects to be considered in respect to the use of multiple regression in studying the relationship between dependent and independent variables in social sciences research. These aspects include types of data, variables selection method, multicollinearity, influence statistics, leverage, Cook's distance, standardized regression coefficient, coefficient of determination and violation of assumptions.

REFERENCES


