Multiple regression modelling

Contents

Introduction

Simple regression and correlation relate to just two variables, and we concentrated on just linear relationships.

One of the measures we derive from a simple regression equation is the coefficient of determination, r^2, which can also be obtained by finding the correlation coefficient, r, between two variables, then squaring it. You will have found that software output for simple regression often denotes r^2 by R-squared, or R^2. Despite the different notation, the concepts are the same. It is a very useful result since it enables us to provide a quantitative measure of the goodness of fit: 'the proportion of the variability in the dependent Y variable explained by or attributed to variation in the independent X variable'.

R^2 (or r^2) provides us with a good starting point for multiple regression because the obvious question is: Can we improve on R^2? The answer is, generally, Yes. In general, the higher is R^2, and the closer it is to 1.00 or 100%, the better the fit and the more confidence we have in using the regression model generated.

If we had to be limited to simple regression to develop a regression model with Y as the dependent variable, but we had several X variables to choose from, then, if we wanted the best model, it makes sense to use that X variable which produces the highest R^2. The value of multiple regression, however, is that we can add other X variables to our model, and, in general, each time an X variable is added, the R^2 increases.

Multiple regression analysis, then, allows for more than one independent variable to be included in determining the dependent variable. It allows for relationships between many (independent) variables and the dependent variable of interest. Thus, we have several X variables, but only one Y variable.

For example, it may be of interest to model the level of sales of a given product for an organisation. This is the dependent Y variable. There are many candidate variables that could be considered as independent X variables: variables such as price, amount spent on advertising, amount spent on other forms of promotion, distribution channels used, income of the target segment, interest rates, other economic factors and level of competition. Multiple regression can determine which of the variables are important and which are unimportant in explaining variation in sales. Further, it provides a quantitative representation of the effect of each variable on the dependent variable. Management can use this information to determine strategies and predict and control sales results, with subsequent impact on budgets and efficiency of resource use.

Multiple regression analysis, properly used, is extraordinarily powerful and one of the world's most useful research tools. It has several major uses:

- explanation of variation in the dependent, Y, variable

- interpretation of coefficients for decision-making purposes regarding *Y*
- prediction of the value of *Y* on the basis of values of the *X* variables
- control of the *Y* variable by varying the *X* variables.

Objectives

- explain the uses and value of multiple regression
- use correlation analysis and scatter plots to help determine suitable X variables for a multiple regression model
- conduct multiple regression analysis on relevant variables and generate appropriate multiple regression models
- conduct inferential analysis on relevant sections of the model
- understand the assumptions underlying the regression model
- incorporate categorical variables into the regression model
- allow for a non-linear effect in a regression model
- explain the role of automated methods in regression modelling
- complete some simple diagnostic checking on the validity of the results and any assumptions and assess the overall use of a particular model.

Developing the multiple regression model

The simple regression model which relates two variables *Y* (dependent) and *X* (independent) may prove insufficient to adequately explain real-world phenomena.

For example, there are many factors which influence sales: the level of price, the type of promotion, the demographics of the target market, etc. If we wish to build a model to adequately explain the level of sales, we need to explicitly include variables such as these in the regression model. Thus, we will need more than one independent (explanatory, predictor) variable. The number of dependent variables remains at one, however. We can expand the simple regression analysis into a multiple regression analysis.

Typically, a linear model is considered as follows:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \varepsilon_i$$

where β_0 = *Y* intercept

β_p = slope of *Y* with variable X_p holding variables $X_1, X_2, X_3, \ldots X_{p-1}$ constant.

To estimate this model we need to gather data on *Y* and each of the *X* variables jointly.

Software packages can produce a large range of regression output, including the regression equation, the individual coefficients, *p*-values, calculated residuals and residual plots. See exhibit 1 for an example.

Exhibit 1: Multiple regression output for Houseprice

Analysis of Variance, ANOVA							
	Degrees Freedom, df	Sum of Squares, SS	Mean Square, MS	F-Ratio	p-Value		
Regression	2	394875629631	197437814815	21.486	0.00000		
Error	27	248102037036	9188964335				
Total	29	642977666667					
Regression Equation Results							
Dependent Variable, Y: Price ($)							
Price ($) = 263948.172 + 288.421 Size (m2) -204.341 Distance (m)							
Indep. X Variables	Coefficient	Standard Error	t-Statistic	p-Value	95% Conf. Lower	95% Conf. Upper	VIF
Intercept	263948.172	80513.809	3.2783	0.00287	98747.484	429148.86	
Size (m2)	288.421	117.285	2.4591	0.02062	47.772	529.07	1.004
Distance (m)	-204.341	32.831	-6.2240	0.00000	-271.704	-136.977	1.004
R-Squared	61.41%						
Multiple R	0.7837						
Adj. R-Squared	58.56%						
Standard Error of Estimate	95859.086						
Durbin-Watson	1.051						
Number of Observations	30						

The interpretation of the results of a multiple regression is similar to simple regression but differ in the following way: the coefficients b_1, b_2, etc. (which are the estimated slope coefficients from the model) are interpreted individually. For example, we say that b_1 represent the estimated average effect on the dependent variable of changing the X_1 independent variable by one unit, provided all of the other independent variable values remain unchanged. They are sometimes called conditional coefficients. From exhibit 1, 288.421 represents that for every extra 1 metre square, the estimated average effect on house prices is $288, provided distance remains unchanged.

Coefficients of multiple determination

In multiple regression analysis we can calculate measures which indicate the explanatory power of the regression model. These measures relate to the proportion of variation in the dependent variable that is explained by the set of explanatory variables as a group.

The coefficient of multiple determination, R^2, can be used to measure the explanatory power of the multiple regression model (similar to simple regression).

An area of caution with multiple regression is that adding new variables into the model will never decrease the explanatory power of the model, even if the added variables are intrinsically unrelated to the dependent variable. Thus, the R^2 could be artificially increased by adding insignificant variables into the regression model. Thus, it can be misleading to compare R^2 for regressions for the same dependent variable where there are different numbers of independent variables.

A measure that seeks to overcome this drawback with R^2 is the adjusted R^2. It includes an adjustment for the number of explanatory variables in the model. It can decrease if the added explanatory variables do not add *significantly* to the explanatory power of the model. It can be used to compare regressions for the same dependent variable with different numbers of independent variables. For example, when adding an explanatory variable into the regression model and re-estimating the equation, a decrease in the adjusted R^2 suggests that the added explanatory variable does not add significantly to the explanatory power of the model. We might consider dropping the variable from the model and using the previous model.

The adjusted R^2 and the R^2 are both provided in the software output. Examine exhibit 1. The information is contained at the bottom of the output. In this case the R^2 indicates that 61.41% of the variation in Price (Y) can be explained by the variation in the two independent variables Size and Distance (X_1,X_2).

The adjusted R^2 is 58.56%. This says that, adjusted for the number of independent variables, 58.56% of the variation in Price is explained by the explanatory variables.

The main use of the R^2 adjusted is to compare models for the same dependent variable with different numbers of explanatory variables. However, do not rely on either R^2 or the adjusted R^2 as the sole means of deciding whether or not a model is a good one, or a better one. Thus, do not use one at the exclusion of the other. While there is the caution about using R^2 as it can provide an over-optimistic view of the level of explanation, it is still a valid observation about a model. In particular, use the two R^2 results in conjunction with the other regression results to evaluate the suitability of a model and for comparing models.

Statistics authors also advise that neither R^2 nor the adjusted R^2 are appropriate for comparing models where the dependent variable are different or where the sample sizes are different.

Preliminary correlation and scatter diagram analysis

You need a starting point when trying to determine an appropriate regression model. Hence, you need to undertake some prior-model analysis. The two most useful are correlation analysis for detecting *linear* relationships between variables and scatter diagram analysis to confirm those linear relationships and to help detect any *non-linear* (curved) relationships.

The correlation matrix and scatter plot approach serves two main functions.

1 It can detect strong (or strong enough) linear associations between the dependent and independent variables, that is, between Y and any X variable. Thus, it helps detect which X variables are important in explaining the variation in the dependent variable and which should be considered for inclusion in the model. This gives us a list of *potential* X variables.

2 It can be also be used to check if there are strong correlations between the independent X variables in the *potential* list. Including two X variables that are highly correlated in a multiple regression model may cause problems in estimation and can lead to misleading results. (See collinearity later.) Thus, we use correlation analysis to detect potential collinearity problems, enabling us to exclude one or more variables from the *potential* X list. Low correlations among potential independent variables are acceptable.

Correlation analysis is a quick way of detecting *linear* relationships between pairs of variables. However, sometimes there is a very important and useful *non-linear* relationship between Y and an X variable. Scatter diagram analysis is the appropriate method for that (and for confirming the linear relationships from the correlation analysis). It is often possible to allow for non-linear relationships in multiple regression, leading to a vastly improved model. (See later.)

Pre- and post-model analysis

The correlation and scatter diagram analysis are the main pre-model building options. They enable us to identify a set of variables as a starting point for developing a suitable regression model. The model building itself is based largely on trial and error and experience. Each time a model is generated it has to be evaluated in a number of different ways. These post-model tools including the overall F-test, checking R^2 and adjusted R^2, potential collinearity issues, significance of each variable in the model and residual analysis. These post-model tools are discussed below.

Test for the significance of the overall multiple regression model

When a multiple regression model is generated, an obvious question is ‘Is at least one of our variables significant in this model?’ Clearly, if the answer is no, then we need to generate another model. If the answer is yes, then we can proceed to analyse the remaining features of the model. Testing for the significance of the overall model, called the *overall F test*, is an example of a hypothesis test.

The null hypothesis can be read as ‘none of the *X* variables are important in this relationship’ while the alternative could read ‘at least one *X* variable is important’.

As a general rule use the *p*-value (in the ANOVA table. See exhibit 1) for choosing between these two hypotheses. If the *p*-value is less than your level of significance, α, (usually 5%), reject H_0 (in this instance your *F*-ratio would be relatively large). If the *p*-value is greater than α, do not reject H_0 (in this instance your *F*-ratio would be relatively small). As the value of the *F*-ratio cannot be interpreted directly, do not use it as the main guide—the *p*-value is directly comparable to α.

APPLICATION

Suppose that a large consumer products company wants to measure the effectiveness of different types of advertising media. Specifically, two types of advertising media were to be considered: radio and television advertising, and newspaper advertising. A sample of 22 cities with approximately equal populations were selected. Each city was allocated a certain advertising expenditure for both radio/TV and newspaper. The sales of the product ('000s) during the test month were also recorded.

Output for multiple regression results are given below. (Note that the correlation matrices would be generated first in a typical modelling situation.)

Analysis of Variance, ANOVA							
	Degrees Freedom, *df*	Sum of Squares, *SS*	Mean Square, *MS*	*F*-Ratio	*p*-Value		
Regression	2	2028032.690	1014016.345	40.158	0.00000		
Error	19	479759.901	25250.521				
Total	21	2507792.591					
Regression Equation Results							
Dependent Variable, *Y*: Sales							
Sales = 156.430 + 13.081 Radio + 16.795 Newspaper							
Indep. *X* Variables	Coefficient	Standard Error	*t* Statistic	*p*-Value	95% Conf. Lower	95% Conf. Upper	VIF
Intercept	156.43	126.758	1.2341	0.23222	-108.877	421.738	
Radio	13.081	1.759	7.4349	0.00000	9.398	16.763	1.009
Newspaper	16.795	2.963	5.6676	0.00002	10.593	22.998	1.009
R-squared	80.87%						
Multiple *R*	0.8993						
Adj. *R*-sq	78.86%						
St Error	158.904						
Observations	22						

We need to evaluate the worthiness of this model. The recommended starting point is the F-test, using the *p*-value and *F*-Ratio given above. From the ANOVA table, we see that the *p*-value is very low (zero to five decimal places) meaning the *F*-Ratio of 40.158 is relatively high. This tells us at least one variable is significant in the model. (More on testing individual variables later.)

Then consider the goodness of fit. The R^2 is 80.87%, thus we estimate that about 81% of variation in Sales is explained by or attributable to variation in the two X variables, Radio and Newspaper. About 19% remains unexplained. An R^2 of about 81% is a good result, implying a high degree of fit and giving us increased confidence in using the model, should we find the model is acceptable overall.

The VIFs at the right of the exhibit will be explained later, but if any are 5 or more, there is the possibility of collinearity, meaning that, in effect, the same variable has been included in the model twice, and that aspects of the rest of the model may be adversely affected. In particular, the coefficients of one or more X variables may be incorrect or nonsensical or some variable that should be significant in the model aren't while some are shown as significant, when they are not expected to be.

Whether or not the VIFs are high or low, the coefficients and significance of each variable must be checked.

The estimated equation from the results is:

Sales = 156.430 + 13.081 (Radio/TV) + 16.795 (Newspaper)

Do the coefficients for X_1 and X_2 in this regression model conform to expectations, and do they make reasonable sense? In this case, they do, as increases in advertising levels from either medium should increase sales. We expected the signs to be positive and they were. (Had we generated the correlation matrices, the individual correlation coefficients could be checked to confirm this.)

In a section to follow, we will show how to check for the significance of each variable. In this example, both Radio and Newspaper are significant in the model (*p*-values less than 5%). Hence, we can proceed to interpret the coefficients in detail.

The individual coefficients can be interpreted as follows.

The intercept of 156.430 is the approximate average sales if there was a zero level of both radio and newspaper advertising. This seems plausible, particularly since the data set show that both Radio and Newspaper showed zero for some observations. (However, an examination of the data reveals that there are no cities where the levels of advertising both approached zero so caution in interpretation is recommended. Often b_0 does not have a useful/plausible practical meaning.) In this case, the point estimate of the average level of sales would be 156.4304 ($'000s) if there were no radio or newspaper advertising.

The slope coefficient for radio/TV is 13.0807. This suggests that, *on average*, an additional unit of radio/TV advertising ($'000s) will increase sales by 13.0807 units ($13,087), assuming that *everything else was unchanged*— that is, in this case, provided the level of newspaper advertising was held *constant*.

The slope coefficient for newspaper is 16.7953. This suggests that, *on average*, an additional unit of newspaper advertising ($'000s) will increase sales by 16.7953 units ($16,795.3), assuming that *everything else was unchanged*—that is, provided the level of radio/TV advertising was held *constant*.

In this example, we can compare the effects of the two variables because they are measured in the same units ($'000s). However, in regression models where the variables are measured in different units, you cannot use the coefficients for direct comparison. For this model it appears as if, dollar for dollar, newspaper advertising has a greater impact on sales than does radio/TV advertising.

The multiple regression equation can be used in much the same way as the simple regression equation. In this case, management can evaluate, in terms of sales, the effectiveness of a particular promotional strategy. Suppose management was considering two strategies. The first involved radio/TV advertising of 50 ($50,000) and newspaper advertising of 20 ($20,000). The second strategy was radio/TV of 30 ($30,000) and newspaper advertising of 40 ($40,000). Which strategy should management prefer, in terms of larger sales? (Note that the total cost of each strategy is the same.)

To evaluate the strategy problem, we need to determine the levels of sales that are predicted in both cases by the regression model. This is a simple case of substitution into the equation.

The equation is:

Sales = 156.430 + 13.081 (Radio/TV) + 16.795 (Newspaper)

For strategy one:

Sales (S1) = 156.430 + 13.081 (50) + 16.795 (20)

= 1146.37 ($'000s)

For strategy two:

Sales (S2) = 156.430 + 13.081 (30) + 16.795 (40)

= 1220.66 ($'000s)

In terms of sales, the second strategy is preferred. In this example, the independent variables were both promotional variables and were measured on the same scale ($'000s). The total cost of the different strategies was easily compared and constrained to be equal. In other circumstances where the independent variables are not similar variables, it would be wise to evaluate the costs and revenues implied by each strategy before deciding which strategy is preferred.

Residual analysis

Just as in the simple regression case, in multiple regression it is important to check the validity of the assumptions that underlie the model. The assumptions of the multiple regression analysis are similar to that of simple regression. Residual analysis is one way in which the validity of the regression assumptions can be checked.

The important difference here is that residual plots can be formed against *all* of the explanatory variables in the model, against time for time series data or even against the *Y* predictions in the model. As in the simple regression case, a non-random scatter may indicate that there are assumption violations and that the regression results are invalid. You should check for a random scatter in all of the relevant plots. (For dummy explanatory variables the scatter will not be random due to the nature of the variable. These variables will be discussed later.)

Inferences concerning the population regression coefficients

As with simple regression, in multiple regression we are primarily interested in the population relationship between the dependent variable and the explanatory variables. The sample regression coefficients represent the estimated results for the sample. We would like to generalise these results or link them to the population regression equation. We do this via inferential procedures. The main tests that are performed are tests of significance. These tests ascertain if individual explanatory variables are important in helping explain the variation in the dependent variable, when the effects of the other variables are held constant. An X variable is said to be related to Y if there is a tendency for Y to increase as X increases meaning there is a positive slope in the relationship, or for Y to decrease as X increases, meaning there is a negative slope. If there is no tendency for Y to increase or decrease as X increases, then we say there is a random horizontal pattern with zero slope.

Thus, zero slope means no relationship between the X variable and Y and that X has no explanatory or predictive power for Y.

The hypothesis test procedure sets the hypothesis as:

H_0: $\beta_k = 0$

H_1: $\beta_k \neq 0$

Where β_k refers to the regression coefficient for variable X_k.

H_0 hypothesises that the chosen variable X_k is not significant ('has a slope of zero'; 'has no explanatory or predictive power'), while H_1 indicates otherwise.

Computer output generally assumes the above hypothesis test and calculates the required *t*-statistics and *p*-values for each coefficient. While a rough rule of thumb of ±2 can be used for the *t*-statistic as to whether or not a variable is significant, the easiest and more accurate approach is to use the *p*-values. Thus, a small *p*-value (that is, less than your level of significance, say, α = 5%) and correspondingly large *t*-statistic (greater than about 2 in magnitude) indicate significance , while a large *p*-value and corresponding *t*-statistic that is small in magnitude indicate not significance.

The other inferential technique that can be applied to the coefficients is confidence intervals. Confidence interval estimates for coefficients are determined in a way similar to the simple regression case. If the confidence interval for a coefficient straddles zero (that is, the lower limit is negative and the upper is positive) it seems that the equivalent population parameter could be zero (meaning that H_0 in the test given above would not be rejected). If the confidence interval does not straddle zero (that is, the lower and upper limits are both positive or both negative) then it seems that the equivalent population parameter is not zero (meaning that H_0 in the test given above would be rejected). If we conclude that a coefficient is not zero, the obvious question is 'what is its value?': the confidence interval gives us the range in which we are very confident (for example, 95%)

APPLICATION

A mail-order catalogue business selling personal computer supplies, software and hardware. It maintains a centralised warehouse for the distribution of products ordered. Management is currently examining the process of distribution from the warehouse and is interested in the factors that influence warehouse distribution costs. Currently a small handling fee is added to the order, regardless of the amount of the order. Over the past 24 months data have been collected which indicate the warehouse distribution costs, the sales and the number of orders received. Management wishes to examine an appropriate regression model to predict the distribution Cost (Y) based on the predictor/explanatory X variables, Sales and Orders.

Our first step is to generate a correlation matrix for the variables. The correlation matrix for the data appears below.

Correlation Coefficient: *r* - Combined matrices			
	Cost	Sales	Orders
Cost: *r*	1.000	0.842	0.919
t-stat	–	7.324	10.918
p-val	–	0.00000	0.00000
Sales: *r*	0.842	1.000	0.800
t-stat	7.324	–	6.260
p-val	0.00000	–	0.00000
Orders: r	0.919	0.800	1.000
t-stat	10.918	6.260	–
p-val	0.00000	0.00000	–

The correlation matrix shows that both explanatory variables seem strongly linearly related to the dependent variable, Cost. (Note the high *t*-statistics and the low *p*-values.) There is also a healthy correlation of 0.8 between sales and orders which may affect the regression results, due to collinearity (sometimes known as multi-collinearity). This concept is explained later, but largely it means that if two *X* variables are strongly related to each other, the resulting regression model may be misleading or erroneous. Despite this concern, we will try both variables as explanatory variables in a multiple regression model. (If we had low correlation between Cost and one of the X variables we should use scatter diagrams to see if there was a potential non-linear/curved relationship.)

The multiple regression results are shown below. From the output, the estimated equation is:

$$\text{Cost} = -2.728 + 0.047\ (\text{Sales}) + 0.012\ (\text{Orders})$$

The signs on both Sales and Orders concur with logic. They are both positively related to Costs and the signs for both variables are positive, as the r values indicate should be the case.

This equation can be used for interpretive and explanation purposes and prediction of Cost (given sales and order values), provided that the post-model diagnostic checks and inferential procedures indicate otherwise.

We can only apply the equation above if the inferential procedures and diagnostic checks indicate model validity.

Analysis of Variance, ANOVA							
	Degrees Freedom, *df*	Sum of Squares, *SS*	Mean Square, *MS*	*F*-Ratio	*p*-Value		
Regression	2	3368.087	1684.044	74.134	0.00000		
Error	21	477.043	22.716				
Total	23	3845.130					
Regression Equation Results							
Dependent Variable, *Y*: Cost							
Cost = -2.728 + 0.047 Sales + 0.012 Orders							
Indep. *X* Variables	Coefficient	Standard Error	*t* Statistic	*p*-Value	95% Conf. Lower	95% Conf. Upper	VIF
Intercept	-2.728	6.158	-0.4430	0.66226	-15.534	10.078	
Sales	0.047	0.02	2.3177	0.03064	0.005	0.089	2.781
Orders	0.012	0.002	5.3131	0.00003	0.007	0.017	2.781
R-squared	87.59%						
Multiple *R*	0.9359						
Adj. *R*-squared	86.41%						
Standard Error of Estimate	4.766						
Durbin-Watson	2.258						
Number of Observations	24						

The inferential test for the overall usefulness of model is the *F-Test.* The observed value of the *F-statistic* (or *F-Ratio*) is 74.134 and the *p*-value is very small, at zero to five decimal places. This low *p*-value and high *F*-statistic indicate support for the population regression model having some explanatory power for the dependent variable—that is, at least one of sales and orders is useful in explaining variation in *Y*.

The model has reasonable explanatory power (R^2 = 0.8759, R^2 (adjusted) = 0.8641). A check of the coefficients has shown they make logical sense.

The inferential procedures for the regression coefficients are based on the *t*-statistics. Since both *t*-statistics for sales and orders are, in absolute terms, larger than the rule of thumb value of 2, and the *p*-values are less than 5% or 0.05 (0.030644 and approximately 0.00003 respectively), then each variable separately (and holding the influence of the other constant) seems to be an important determinant of Cost for the population regression model. In both cases H_0 (that a variable is not significant in the model) is rejected.

The final check on the model is through residual analysis. In particular, we need to check for patterns indicating violations of any of our regression assumptions. We want to check the relevant residual calculations and plots. The most efficient means of doing this is via a range of residual plots (as seen with simple regression).

All of the above inferential procedures (assuming residual plot checks are OK) confirm that the model seems an adequate explanatory model for Cost. Since we do not know the population model we will use the sample regression results as estimators of the equivalent population parameters. Our sample regression equation is:

Cost = – 2.728 + 0.047 (Sales) + 0.012 (Orders)

In this model we suggest that:

- the intercept term of -2.728 has no useful meaning since it implies costs could be negative when there are zero sales and zero orders
- a one unit change in sales will lead, on average, to a 0.0471 unit change in Cost (all other factors held constant)
- a one unit change in orders will lead, on average, to a 0.012 unit change in Cost (all other factors held constant)
- this model can then be used for prediction of distribution costs given sales and orders.

Note that we cannot compare the size of the regression coefficients meaningfully as they are in different size dimensions ($'000s for Sales v Number of orders for Orders).

Dummy variables

A dummy variable is normally a coded categorical with just two values, 0 or 1. With the variable Gender, it can be recorded as 0 for Males and 1 for Females. This is automatically a dummy variable, as there are only two possible outcomes. However, say an organisation's employees work in three different departments: Administration, Production and Distribution. There are three possible outcomes. We could break this into two outcomes: 'Administration' or 'Not Administration', 'Production and Not Production' and

‘Distribution and Not Distribution’. Those columns could then be legitimately included in certain statistical calculations, and in regression in particular.

The regression models used up until now have used numerical variables for both the dependent variable and independent variables. The dependent variable must be numerical for the regression model to be valid. However, in many business situations, non-numerical variables are important explanatory variables.

If possible, categorical (or qualitative) information should be incorporated into a regression model to enhance the predictive ability. For example, gender or department may be important determinants of the productivity level of employees or job satisfaction.

Or, whether or not a salesperson has undertaken a special training course, may be an important determinant of the salesperson’s sales performance. To do this we use dummy variables.

The interpretation of the dummy variables in correlation and regression is in many ways identical to the numeric independent variables. The *t*-statistic and *p*-value for the dummy variable is used (as in the previous sections of this topic) for testing for correlation, and for testing for significance in a regression equation.

The dummy variable can be used to differentiate two groups in terms of the dependent variable values. For example, if we suspect that average wages differ for males and females we could use the regression model to explore this possibility. A hypothesis test procedure to test for the difference in averages between two populations or groups. Regression analysis allows for a similar test but can allow for the effects of other variables on the average difference with the inclusion of selected independent variables in the regression model. For the wages example, we can test for average difference in wages of males and females allowing for the effects of experience, age, etc. This is achieved by including these variables in the regression model as well as the dummy for gender. How do we use the regression model to tell us if there are differences in wages for the gender groups? The *t*-statistic and *p*-value on the dummy variable can be used as a measure of significant differences in wages between the two gender groups. A high *t*-statistic or low *p*-value in the regression model indicates a significant difference between the two groups. Thus, we would conclude that the dummy variable is important in explaining some of the variation in the dependent variable.

Dummy variables can also be used to incorporate seasonal effects into regression models for time series variables.

Collinearity (Multi-collinearity)

You must understand the problem of collinearity in regression model building. This concept is often called multi-collinearity. It refers to situations where two or more *X* variables in a regression model are highly correlated. For example, we know that people’s ‘height’ and ‘weight’ tend to be highly

correlated: the taller you are, the heavier you are likely to be. If we were to include both of these in a regression model, say with pulse rate as the dependent variable, then we are in effect including the 'same' variable twice. The effects on the regression model include nonsensical coefficients, coefficients with clearly the wrong sign (for example, a negative coefficient for weight, indicating that the heavier you are, the lower the pulse rate, when in fact it should be the other way round), and variables that we know are strong predictors of *Y* being shown as insignificant in the model. Such models either should not be used or must be used with extreme caution. It is best to avoid collinearity altogether.

To do so, ensure you do the following:

- In a correlation matrix check for pair-wise correlations between potential independent (explanatory) variables. If potential explanatory variables appear to be correlated, think carefully about including both in a regression model. (There are no hard and fast rules about when two *X* variables are correlated enough to cause collinearity, but a *p*-value of less than 5% or even 10% is one starting point; some authors suggest an r of 0.7 or higher as another.)
- In a regression model, check to see if the coefficients or *p*-values appear to be contrary to what is expected. If so, the problem may be multi-collinearity. For example, if income of employees is the dependent variable, and age is one of several explanatory variables, we expect to see a positive sign for the coefficient of age. If it shows as negative in the model, consider reformulating your model by excluding one or more variables.
- In a regression model, use the VIF results—some authors advise that a VIF of 5 or more is an indicator of possible collinearity.

Allowing for non-linear effects

For all of the preceding analysis we have assumed that the relationship between the dependent variable and independent variable(s) was fundamentally linear. In many cases, logic is used to suggest that the relationship may be non-linear. For example, theory predicts that in production there should be a disproportionate decrease in average costs as quantity produced increases, due to economies of scale and learning effects. Past a certain quantity produced, average costs may fall as output increases but at a smaller rate. This suggests that the relationship between these variables is non-linear and should not be modelled as a linear relationship.

Non-linear relationships between Y and an X variable can appear in many ways, including upward sloping curves, downward sloping curves or U-shapes or inverted U-shapes.

The quadratic regression model

The quadratic regression model includes an X^2 term for a given X variable. For example, we could create a new variable Age^2 (found by squaring the values of the variable Age). Then we can create a regression model including X^2 and X (or just with the X^2 term). If both X and X^2 are included, the VIFs will generally be high, indicating, as we would expect, that the two variables are related but both may be left in the model. The inclusion of an X^2 term can make a dramatic change in the R^2 result, indicating a much better fit overall. While the quadratic term may lead to a much improved degree of fit, a downside is that the coefficients of the X and X^2 terms are generally difficult to interpret.

Using transformation in regression models

If there is a non-linear relationship between two variables, it may be possible to 'straighten it out'. For example, while there may be a curved relationship between *Y* and *X*, there might in fact be a good enough linear relationship between *Y* and $\sqrt{X}$ (square root), or between Y and $1/X$ (reciprocal) or $ln X$ (natural log).

The transformations are made on the data in computer software packages before the regression analysis is conducted. Linear regression is applied to the transformed data in the usual manner.

Note that the interpretation of the results of the transformed regression model is not always similar to the previous sections, and depends on the type of transform chosen: sometimes there is no practical interpretation. Some logarithm transformations enable very useful interpretations of coefficients as percentage changes, rather than absolute changes.

Automated methods in model building

You are probably appreciating that regression model building is time-consuming and requires some skill indeed. Further, the more *X* variables you have in your data set, the more complicated it becomes. For example, with one *Y* variable and two *X* variables there are three possible linear regression models you could derive: '*Y* v X_1', '*Y* v X_2' and '*Y* v X_1 and X_2'. If there are three *X* variables, there are seven possible models (three single *X* variables, three with two *X* variables and one with all three *X* variables). If you have 20 *X* variables, the number of possibilities is huge. How do we construct a suitable, useable regression model from so many variables, let alone find the best?

To speed up the model building process, statisticians have developed automated methods based on regression software performing dozens, hundreds or even thousands of regressions in seconds.

Two methods of interest are the *stepwise method* and the *best subsets method*. A brief summary of each follows.

The stepwise method begins by choosing the best model with one X variable. In the next step it creates a second model using the next most useful X variable. The process continues in this way, adding variables, and even removing them if necessary. The process ceases when the remaining X variables add nothing significant to the model in statistical terms. Thus, the process may begin with fifteen X variables and generate a stepwise model with just six of those variables included.

The best subsets method involves generating dozens, hundreds, even thousands, of regression models. Generally, all single X variable models are calculated first, then all two variable models (with X_1 and X_2, then X_1 and X_3, etc), then all three variable models, etc. The software can generally be instructed to list the best in each group (that is, the best single variable model, the best two variable model, the best three variable model) or the best two in each group, or the best three in each group, etc. The model deemed to be the best in a group is chosen on the basis of a criteria such as R^2 or adjusted R^2.

What you should be aware of is that automated methods are not necessarily the answer to good model building. In particular, there are notable advantages in using the ‘longhand’ (that is, non-automated) method. The main problem of automated methods is that they only permit minimal contribution from the modeller: there is not much room for using experience, judgment or even hunches in developing an appropriate model. Further, they do not necessarily avoid problems like collinearity. Thus, in your own model building you could consider developing your ‘best’ model based on your correlation and scatter plot analysis and subsequent regression analysis. Then use automated methods as a cross-check against your preferred model, and then fine-tune your own model if you discover something of value.

Summary

This topic continued our studies of regression and correlation analysis.

The key features from simple regression also apply to multiple regression, where we introduced models with more than one *X* variable.

We found that inferential work on the regression results was similar to simple regression, and that diagnostic procedures such as residual analysis applied in a similar way.

Multiple regression, however, vastly extends the regression modelling capabilities, including the use of dummy variables to allow the incorporation of categorical variables into the model, and transformed variables as a method for allowing for non-linear effects between pairs of variables. We learnt about some problems to watch out for, in particular, collinearity. We also learnt about automated regression procedures and that while they save a great deal of time in terms of generating ‘good’ models, they don’t necessarily generate the best model.

Estimation of regression models is generally from sample data, hence requiring inferential procedures, including confidence intervals and hypothesis tests for the slope coefficients, and tests for explanatory power of the model. The main statistics for inferential procedures are the *t* and *F-statistic*s. Key results from any regression model include the R^2 and R^2 (adjusted) but we learnt not to rely just on the former and to use the adjusted option for helping to choose between different regression models with the same dependent variable but a different number of independent variables.

There are many other features of multiple regression modelling we have not covered, such as how to allow for the effect of individual observations.

Further resources

Black, K 2008, *Business statistics for contemporary decision making*, 5th edn, Wiley, NJ.

Anderson, DR, Sweeney, DJ & Williams, TA 2008, *Statistics for business and economics*, 10th edn, South-Western Thomson Learning, Cincinnati.

Selvanathan, A, Selvanathan, S, Keller, G & Warrack, B 2006, *Australian business statistics*, 4th edn, Nelson Thomson Learning, Melbourne.