

Financial Modelling

**By Franco Azzopardi, MSc student
University of Leicester, 2004**

Contents	<i>Page</i>
Introduction	3
Correlation between variables	4
Regression line	6
Standard error of the regression line	15
Standard error of the coefficients	16
Testing the models – hypothesis testing	17
Testing the significance of coefficients	20
Goodness of fit of the regression line	22
Conclusions	27
References	28

Introduction

Business research and analysis is often concerned with economic relationships and functions between variables. This emanates from the need for a better understanding of economic events, and for policies that are conducive to better implementation of operational, tactical and strategic objectives of the concern. The improved understanding comes about by means of a closer study and more thorough appreciation of the many complex relationships that are entwined within the economic environment. To this end, the analyst has to be in possession of better statistics and improved ways of analysing them, drawing on the quantitative skills of the mathematician, statistician and econometrician, rather than on the traditional belief that just good sense and a sound judgement of affairs is enough.

In an effort to understand the volatility of sales, Zoom Cars has thought of determining whether sales is related in some way to other variables. The analyst has identified and collected data on various variables, and probably measured the degree of association between pairs of variables. This '*correlation coefficient*' is a unit-free measure of the strength and direction of a linear relationship between any two variables, but only a statistical measure of association. There is no inference of '*causality*' in the statistic. This means that cause and effect may be present but correlation does not prove cause. The analyst has established that there exists a functional relationship between sales and prices, levels of advertising and favourable reviews in the motoring press. A priori reasoning develops the theory of the causal relationship between these pairs of variables, concluding that sales is a variable that depends to some extent or another on the other variables. Hence, sales is said to be the dependent variable and the others, the independent variables. From this reasoning and from the data collected, '*regression analysis*' is used to construct a model that reflects the hypothesised relationship and then test the model statistically. One of the main uses of regression analysis is as a prediction tool. The analyst of

Zoom Cars has come up with two models. The scope of this report is to address the concerns of the company by analysing the findings. The approach that will be applied in the paragraphs that follow will splice the theory of regression analysis with the practical situation and analysis of Zoom Cars, in stages. This approach is directed at giving the reader an interpretation of the findings with reference to the theoretical principles involved.

Correlation between variables

It is said that knowledge is power and effective business management must harness information through proper and robust information systems. These information systems should provide, amongst other things, statistical data about variables that may have an effect on the operations, marketing and selling functions of the business concern. The strength of the relationship between pairs of variables is evaluated through the analysis of their relevant data and through the application of statistical techniques such as the Pearson Correlation Coefficient. This correlation coefficient is named after Karl Pearson (1857-1936) who developed the current version in 1896 in England¹. The coefficient, or number, that represents the correlation is always between +1.00 and -1.00. A perfect positive correlation (+1.00) indicates that every time one variable increases in size, the other variable also increases in an exact linear fashion. A perfect negative correlation (-1.00) indicates that every time one variable increases in size the other variable decreases in an exact linear fashion. A correlation coefficient of 0.00 means that there is no relationship between variables. The Pearson Correlation Coefficient is applicable only if both variables under consideration are measured on an interval scale. It is normally denoted by:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} \quad (\text{equation 1})$$

where r is the correlation coefficient,

¹ Kotz & Johnson, 1982, p.199

x is one variable and \bar{x} is the arithmetic mean of the same variable.

Similarly, y is another variable and \bar{y} is the mean of that variable.

The arithmetic mean is the sum of the data of the variable of specific interest divided by the number of observations.

The usefulness of the coefficient is based on the size and significance of r . As discussed earlier, the correlation is perfect if r is $+1.00$ or -1.00 . One of the main purposes of correlation is the prediction of one variable from the value of another variable for a given subject. For predictive purposes, absolute r values lower than about 0.7 may produce unacceptably large errors in an individual prediction, especially if standard deviations of either or both variables are large. As a general rule, an absolute r value of 0.5 to 0.7 is considered low, 0.7 to 0.8 is moderate, and 0.9 and higher is good for predicting y from x values². If two variables are correlated at $r=0.8$, then they have 64% (explained by $r^2 = 0.8^2$) common variance. This means 64% of the variability in y can be explained by variance in the x variable. The remaining unexplained 36% of the variance is responsible for error when predicting y from x . When r is evaluated, the number of pairs of values from which the coefficient was calculated is critical for determining the odds that the relationship could have happened by chance. Using the t -distribution statistics, one can find the significance of r . When the r value is found to be significant, the cause of the correlation still cannot be determined from the correlation data alone. However, reasoned logic may point to a probable cause.

The researchers of Zoom Cars have identified a number of variables that they consider to be determinant factors in shaping a model that explains the volatility of sales. Most probably, they would have collated data about various economic and other issues that may have logically been fuelling the said volatility. Apart from the price charged, advertising expenditure and favourable reviews in the motoring press, the research probably included also financing terms allowed by the company itself or by intermediaries. The cross selling drives by financial intermediaries promoting equity release loans for example, might be a force behind the fluctuations in the sales. Monetary policy and international events might be telegraphing messages of more

² William J Vincent, EdD, *Statistics in Kinesiology*, 1995, p.97

stability or otherwise. Political events in the country could also be having their effects felt. The introduction of stricter legislation and regulation on vehicle roadworthiness tests might be pushing customers to scrap old cars. That said, in Malta there has been a dramatic effect on sales of new cars because of the mushrooming business of second hand car imports from Japan at ridiculous prices. Low to medium income earners can easily afford super-cars and this is scarring the business of new cars. Another variable that could have had a correlation with sales is insurance costs. Insurance companies and associations of such companies have a strong effect on the demographic pattern of car sales in that heftier premiums are burdening youngsters in such a way that they cannot afford to keep a new car, and therefore they opt for second hand cars. The same applies to medium and low-income earners. This brings up another potential variable that may result in some strength in the correlation with sales, and that is the average household income and the disposable income. The varying degrees of propensities to save and to consume and the effect of attractive pension schemes, bond issues and other saving schemes could also be alternatives. All said, these potential variables were probably ranked in order of the term effect and some of them would be classified as having an effect on sales only in the longer term. Since the objective of the company was to analyse the volatility of sales, these said variables affecting sales over longer terms were probably short-listed.

Regression Line (line of best fit)

A visual description of a correlation coefficient between two variables may be presented as a scatter plot. This usually involves scaling the independent variable along the x-axis and the dependent variable along the y-axis. Graphing the data will yield information about shape and spread of the data.

Regression analysis is a method that attempts to establish a linear relationship between two or more variables. In the case where there are only two variables, the technique is called simple or bivariate regression. However it should be noted that in most business situations as in the case of Zoom Cars, there would be one dependent and a number of independent variables. The dependent variable in our case is the sales and the independent variables are the price, advertising and the favourable

reviews, supplemented also by dummy variables which will be explained further on. Therefore the researcher would have to use multiple regression techniques.

Since regression ultimately involves obtaining a linear fit of the points on the scatter plot, in the case of a perfect positively correlated relationship, it would be fairly easy to plot a straight line as shown in diagram 1.

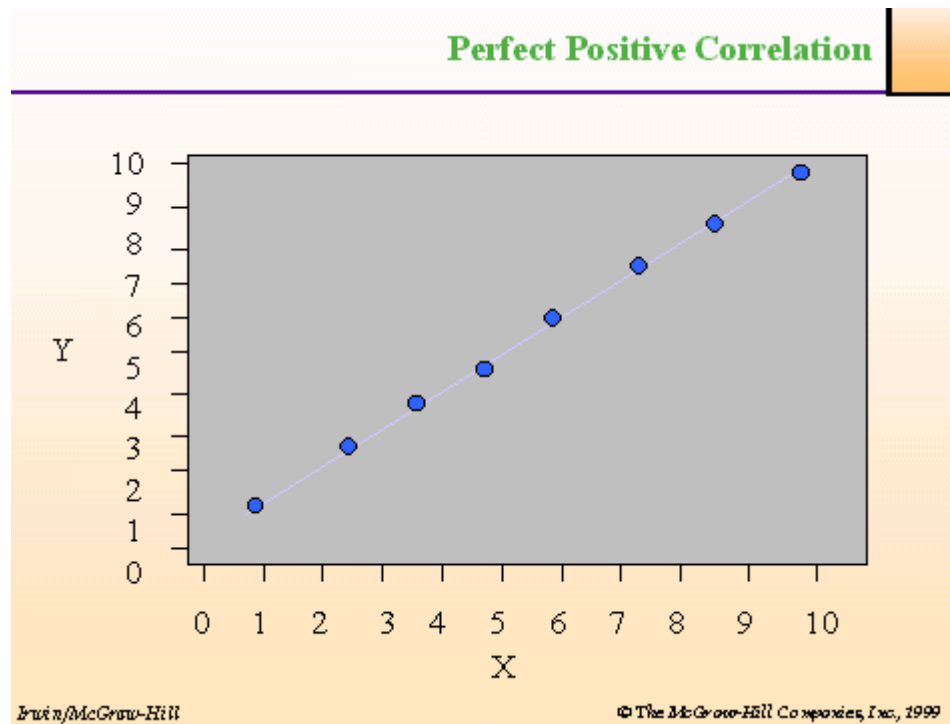


Diagram1 – Scatter plot and best fit line of a perfect positive correlated data³

The same would apply for a perfect negative correlated scatter plot, which would accommodate a perfectly fit straight line passing through the data points but running downward from left to right as shown below in diagram 2.

³ Robert D Mason, Douglas A Lind, William G Marchal, Statistical Techniques in Business and Economics, 10th edition, Student CD-Rom by Irwin/McGraw-Hill

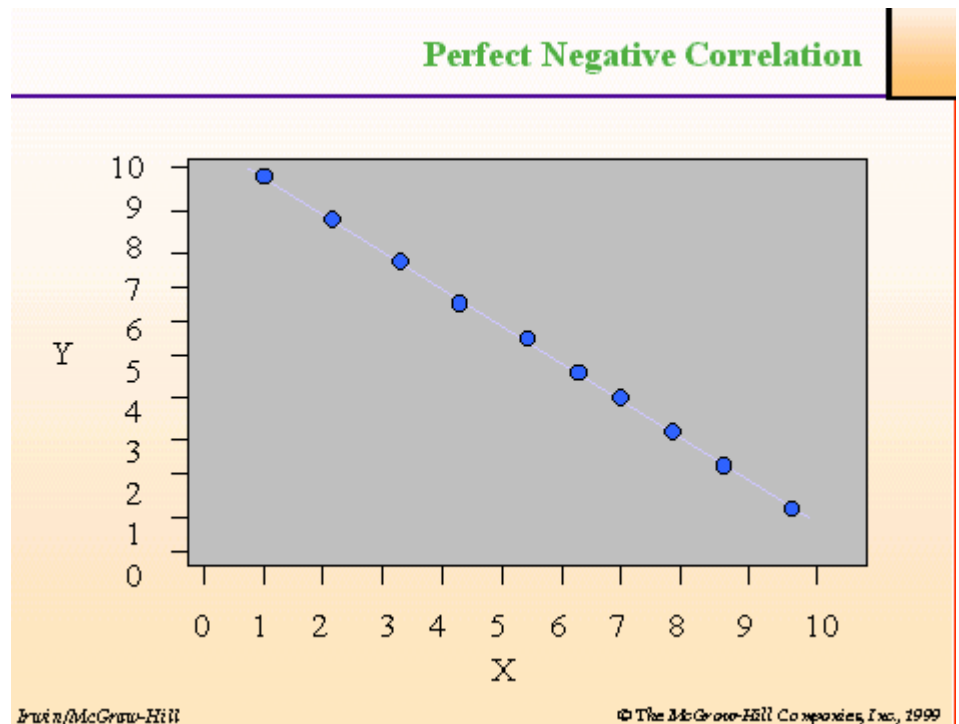


Diagram2 – Scatter plot and best fit line of a perfect negative correlated data⁴

All said, in practical situations, the scatter plot seldom demonstrates a perfect straight line fit passing through all the points. Simple regression techniques suggest an equation that would plot a straight line as a best fit of the data points on the scatter plot. This is demonstrated in diagram 3 below.

The equation of the straight line in the case of a simple or bivariate regression, is shown as:

$$\hat{y} = b_0 + b_1x \quad (\text{equation 2})$$

where \hat{y} = predicted or fitted value of the dependent variable
 b_0 = intercept with y-axis where x is zero
 b_1 = slope coefficient

⁴ Robert D Mason, Douglas A Lind, William G Marchal, Statistical Techniques in Business and Economics, 10th edition, Student CD-Rom by Irwin/McGraw-Hill

x = independent variable

This regression equation is interpreted as saying that the estimated value of the dependent variable \hat{y} , equals a constant amount of b_0 plus b_1 for every unit of the independent variable x .

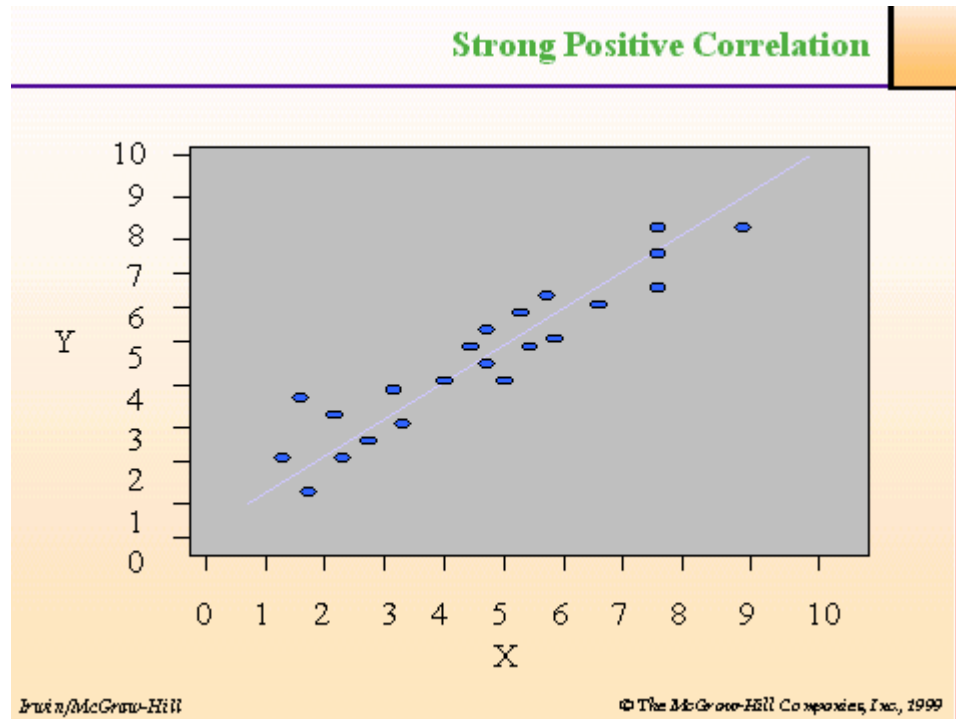


Diagram3 – Scatter plot and best fit line of positive correlated data⁵

In order to determine the equation of a regression line for the sample data, one must calculate the values for b_0 and b_1 . The method of doing this is sometimes called the Ordinary Least Squares estimation (OLS). The estimation method should be such that it is BLUE – Best Linear Unbiased Estimator⁶. This approach gives the following two equations for the regression line:

⁵ Robert D Mason, Douglas A Lind, William G Marchal, Statistical Techniques in Business and Economics, 10th edition, Student CD-Rom by Irwin/McGraw-Hill

⁶ Terry J Watsham & Keith Parramore, Quantitative Methods in Finance, p.190

$$b_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} \quad (\text{Equation 3})$$

$$\text{and } b_0 = \bar{y} - b_1 \bar{x} \quad (\text{Equation 4})$$

where b_1 = slope coefficient (see equation 1)

b_0 = y-intercept (see equation 1)

x = independent variable

\bar{x} = arithmetic mean of the independent variable

y = dependent variable

\bar{y} = arithmetic mean of the dependent variable

If the line representing the equation is superimposed on the scatter plot, it will begin to show how good the equation has resulted. This is demonstrated in diagram 3 on page 9. While the regression equation fits the data well, one can observe the residuals shown by the vertical distance between the line and the individual points. The math of the equation to determine the regression line always brings the summation of the residuals to equal zero, except for rounding errors. An examination of the residuals may shed light on how well the regression line fits the data points and also identify outliers. Outliers must be investigated because they put undue influence on the regression line by pulling it towards them. The technique to do this will be explained further under standard error of the regression on page 15. It is worth noting here that one of the major assumptions underlying regression analysis is that residuals are normally distributed.

As already explained, very rarely do regression analysis involve only two variables. This is because other independent variables taken in conjunction with the original independent variable may improve the regression model for better prediction of the dependent variable. The advantages⁷ of multiple regression over bivariate regression are:

⁷ William J Vincent, Statistics in Kinesiology, p.105

- ✓ Multiple regression usually provides a lower standard error of the estimate and
- ✓ It provides us with information to determine which independent variables contribute to the prediction, and which do not.

In the case of Zoom Cars, the researcher has filtered down to four variables and this warrants an understanding of multivariate regression technique.

The estimated relationship between the dependent variable and the independent variables is given by the equation:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_jx_j \text{ (Equation 5)}$$

where j = number of independent variables

b_j = regression coefficients

b_0 = regression constant

Multiple regression is used to find the most satisfactory solution to the prediction of the dependent variable, which is the solution that produces the lowest standard error of the estimate. However it is more complex conceptually and computationally. The scatter graph and the line of best fit as illustrated on page 9, becomes impossible to illustrate for situations with models containing three, four or even five independent variables. Furthermore, multiple regression model computations are nowadays done using advanced computer programs such as SPSS, BMDP and SAS. The analyst is now mainly specialised in interpreting the computer output.

The similarities between the multiple regression equation 5 and the simple regression equation 2 on page 8, are that they both result in straight lines. The regression coefficients in the multiple regression equation on page 11 represent the partial derivatives of the dependent variable with the appropriate independent variables. The interpretation of the regression constant in the case of multiple regression is more complex than in the case of bivariate regression. For example, in the case of significant constants, it might represent the mean effect on the dependent variable of any independent variable not included in the model.

The analyst of Zoom Cars came up with two regression models as follow:

Model 1

$$\text{LSALES} = -5.226 - 1.039\text{LPRICE} + 2.077\text{LADVER} + 0.300\text{LFAV} \\ - 1.185\text{M7} + 0.926\text{M8} - 0.689\text{M12}$$

Model 2

$$\text{LSALES} = -4.612 - 0.645\text{LPRICE}(-1) + 1.035\text{LADVER}(-1) + 0.866\text{LFAV}(-1) \\ - 1.143\text{M7} + 0.979\text{M8}$$

where

SALES = Sales volume, number of cars sold

PRICE = Average price of Zoom cars in GBP

ADVER = Advertising expenditure by Zoom Cars in GBP

FAV = Favourable reviews in motoring press, number in current month

M7 = Monthly dummy variable, 1 in July otherwise 0

M8 = Monthly dummy variable, 1 in August otherwise 0

M12 = Monthly dummy variable, 1 in December otherwise 0

L represents the natural logarithm of the variable, ln

(-1) indicates that the variable has been lagged one period

The two models seem to differ in format from the multiple regression on page 11. The L-sign in front of all the variables suggest that the logarithmic function of the variables was taken to produce a linear equation from a previously non-linear function. Ignoring for the time being the dummy variables which are explained on page 14 below, the relationship between the non-linear formulation and the linear one is explained as follows:

$$\hat{y} = Ax_1^{b_1} x_2^{b_2} x_3^{b_3} \quad (\text{equation 6})$$

$$\ln \hat{y} = \ln A + b_1 \ln x_1 + b_2 \ln x_2 + b_3 \ln x_3 \quad (\text{equation 7})$$

In this transformation which is referred to as a log-linear model, the coefficient estimates represent elasticities and hence show the percentage or proportionate effect on the dependent variable. This means that in the case of model 1 of Zoom Cars on page 12, a 1% decrease in price of the current month would result in a 1.039% increase in the sales volume of the cars. Similarly, a 1% increase in advertising expenditure in the current month will result in a 2.077% increase in sales volume. Comparing the coefficients of the two models will reveal that the basic difference is the lagging of the statistics by one month in model 2. It transpires from the two models that in the case of the price and advertising expenditure, the effect on the sales is more marked with the current month's data. A priori reasoning supports this, as one would logically expect a higher impact on sales volume if the current month's selling price is reduced rather than if the previous month's selling price had reduced. The same would apply to advertising in that current month's advertising would have a greater effect on sales. Favourable reviews in motoring press would probably build a certain amount of interest in the cars but would not be determining factors to outright purchase decisions by the consumers. The models presented do support this reasoning and this is evident in the higher percentage increase in sales when the previous month had shown favourable reviews.

No information is given in the analysis about the type of vehicles Zoom Cars is trading in. The type of correlation that is being presented seems to suggest that the cars in question are light cars not classy connoisseur vehicles of certain price ranges. In this latter case, there would probably not be that much correlation between sales, price changes and levels of advertising expenditures.

As regards the regression constant, under simple or bivariate regression, the interpretation is simple in that a zero value attached to the single independent variable would produce the point where the regression intersects the y-axis. It has already been explained that in the case of multiple regression, interpretation of the constant is more complex. It is first of all worth noting that the constant in models 1 and 2 on page 12 for Zoom Cars, where it was concluded that the equations are transformations of log-linear models, is the resultant logarithmic function of the

constant in the original non-linear model. This said, it may be interesting to implode the equation of model 1 to its original form as follows:

$$s = \frac{1}{168267.4061} \times \frac{1}{pr^{1.039}} \times adv^{2.077} \times fav^{0.300} \times \frac{1}{10M7^{1.185}} \times 10M8^{0.926} \times \frac{1}{10M12^{0.689}}$$

where s=SALES as on page 12
 pr=PRICE as on page 12
 adv=ADVER as on page 12
 fav=FAV as on page 12
 M7= as on page 12
 M8= as on page 12
 M12= as on page 12

The same type of conversion would apply to Model 2 of Zoom Cars. This type of non-linear formulation is nowadays, through the use of computer software, quite widespread. Otherwise, determining this form of equation from the correlation between data sets without the use of computers would be very complicated.

One may notice that the models resulting from the analysis of Zoom Cars include the qualitative 'dummy' variables M7, M8 and M12. Regression models may sometimes require one or more qualitative variables as opposed to quantitative variables. These variables are coded into a 0 or 1 format so that if they are true, the predicted dependent variable is affected by the coefficient of the dummy variable and if the variable is false, it would have a zero effect on the predicted dependent variable. The inclusion of dummy variables is done to refine the model especially to neutralise the effect of outliers (see page 10). In models 1 and 2 of Zoom Cars, the analyst concluded that in July, sales drop by approximately the same percentage (1.187% in model 1 and 1.143% in model 2). In August, both models acknowledge an increase in sales (by 0.926% in model 1 and 0.979% in model 2). The logical reason behind the inclusion of these dummy variables could be that in July the working society could be ramping up for the summer break and the propensity to buy a car at that time drops. August could be for many, the time to make changes especially that it is the culture of many to take a break at that time of the year, change employment etc. As regards the December seasonal factor included in model 1, the analyst concluded from her

findings that sales are affected negatively by 0.689%. The reasoning here supports the general idea that for the working class society, the festive season would distract consumers from purchasing cars. Model 2 excludes factor.

Standard Error of the Regression Line

We have seen that the regression equation results in a line that best fits the data under examination. However, there will still be residuals (see page 10) shown by the vertical distance between the line and the individual points on the scatter graph on page 9. These residuals can be used to test the estimated equation to determine whether it is a good fit to the data.

A mathematical way of examining the potential error in the regression model is the use of the standard error of the equation which provides a single measure of the model's error.

This is denoted by the following equation:

$$\begin{aligned}
 S &= \sqrt{\frac{RSS}{(n-k)}} \\
 &= \sqrt{\frac{\sum (y - \hat{y})^2}{(n-k)}} \quad (\text{Equation 8})
 \end{aligned}$$

where

RSS is the Residual Sum of Squares found by summing and then squaring the differences between the actual volumes of the dependent variable and the fitted volumes (see \hat{y})

\hat{y} is the fitted volume for the dependent variable found by applying the regression equation to the corresponding independent variable.

n is the number of observations

k is the number of coefficients (in the case of simple regression, there are two coefficients in the equation)

The models that predict the sales of Zoom Cars on page 12 have the following standard errors:

	<u>Model 1</u>	<u>Model 2</u>
$S = \sqrt{\frac{RSS}{(n - k)}}$	$= \sqrt{\frac{0.499}{(36 - 7)}}$	$= \sqrt{\frac{1.465}{(36 - 6)}}$
	= 0.13118	= 0.22098

RSS in the above equations was given by the computer output of Zoom Cars. The value of n , 36, that is the number of observations, is also data that has been provided by the analyst. k is the number of coefficients in the equations, that is 7 in model 1 and 6 in model 2. The calculated standard errors of the models suggest that model 2 gives a higher standard error.

Standard Error of the Coefficients

The next evaluation requires the calculation of the standard errors of the coefficients, which stated simply, are their standard deviations.

The standard error of the intercept is calculated as:

$$SE \text{ of } b_0 = \sqrt{\frac{\sum (y - \hat{y})^2 \sum x^2}{n \sum (x - \bar{x})^2}} \quad (\text{equation 9})$$

The standard error of the slope coefficient is calculated as:

$$SE \text{ of } b_1 = \sqrt{\frac{\sum (y - \hat{y})^2}{\sum (x - \bar{x})^2}} \quad (\text{equation 10})$$

where the symbols are assigned the same meaning as for the previous equations.

In the case of multivariate regression analysis, as has already explained, the computations become highly complex and are therefore normally generated by sophisticated computer applications. Therefore, in the case of multiple regression, the standard errors of the coefficients are given and these allow the calculation of appropriate t-statistics for interpretation purposes. This will be explained on page 19. The standard errors of the coefficients for model 1 and 2 of Zoom Cars have been given as part of the analyst's results.

Testing the models – Hypothesis Testing

One must recall at this stage that the model of the regression line is determined by sample data and not the actual population data. Therefore if a different sample is taken from the same population, a different model would probably result. Thus the

sample coefficients in the regression equation are functions of the particular sample from which they were obtained. The goal of the analyst is to test the coefficients by determining whether the model can add significantly to the explanation of the dependent variable. If the population data were available, would the coefficients of the regression line be significantly different from zero? The reason the analyst asks this question is because zero-coefficients in the regression equation would not contribute anything to the determination of the dependent variable. The analyst normally carries out what is termed the Hypothesis Test. In the case of the simple or bivariate regression, the hypothesis would be as follows:

$$H_0 : B_0 = 0$$

$$H_1 : B_0 \neq 0$$

$$H_0 : B_1 = 0$$

$$H_1 : B_1 \neq 0$$

where H_0 represents the null hypothesis

H_1 represents the alternative hypothesis

B_0 is the population y-intercept

B_1 is the population slope coefficient

One of the basic assumptions underlying regression analysis is that the data have a normal or Gaussian distribution. Therefore the difference between a variable and its mean divided by the estimate of its standard deviation has a t-distribution.

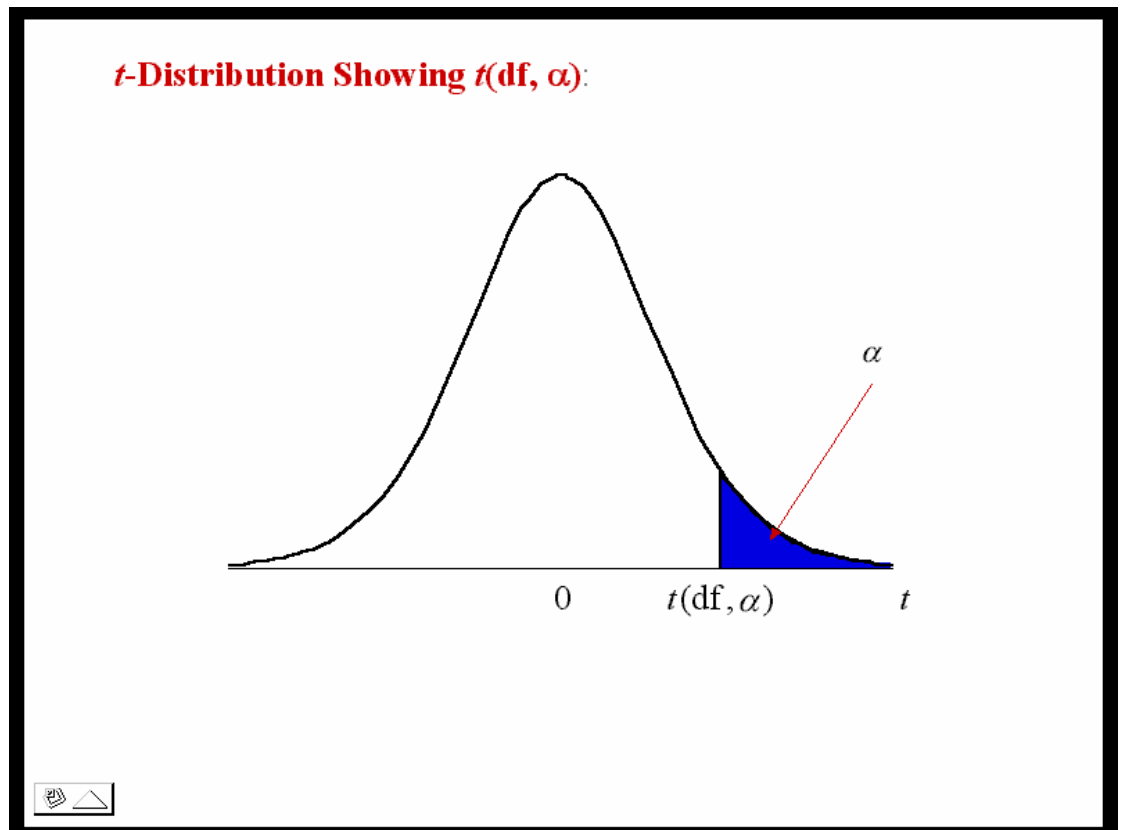


Diagram4 – t-distribution at df degrees of freedom and α level of significance⁸

It therefore transpires that in order to test the above hypotheses one must make use of the t-statistic for the appropriate coefficient. The t-statistics are calculated as follows (slope coefficient taken for the example):

$$t = \frac{b_1}{SE \text{ of } b_1} \quad (\text{equation 10})$$

The value resulting from the calculation needs to be compared with the critical values from the t-distribution tables, with $(n - k)$ degrees of freedom. It is usual to test for statistical significance at the 95% or the 99% levels of confidence. This will result in the 95% or 99% probability that the values of B_0 and B_1 are not due to chance. The regression coefficients are significant if the t statistic resulting from the equation is greater than the value given in the t distribution tables. The level of significance α

⁸ Robert D Mason, Douglas A Lind, William G Marchal, Statistical Techniques in Business and Economics, 10th edition, Student CD-Rom by Irwin/McGraw-Hill

shown in the diagram represents the proportion of the tail of the curve to the remainder of the same curve. Therefore if the test is said to be a one-tailed test, meaning that the expected relationship between the coefficients can be either positive or negative but not both, then this proportion will be 100% less 95%, that is 0.05 where the test is at 95% confidence. If the test is a two-tailed test, as in the above hypothesis test, meaning that the relationship can be both positive and negative, then the proportion will be $\frac{100\% - 95\%}{2}$ that is 0.025. The critical values of the t-distribution are normally symbolised as $t_{\alpha/2, n-k}$ meaning that, the critical values of the t-distribution are two-tailed and at n-k degrees of freedom. With larger datasets, say greater than 30 samples, most researchers will accept a coefficient as significant, if the t-statistic is greater than 2 in absolute value⁹.

Testing the significance of coefficients in multiple regression models

In the case of multiple regression analysis, the standard errors of the coefficients or sometimes the t-statistics are indicated on the computer generated output. In the case where only the standard error of the coefficients is given, the t-statistics for each independent variable will have to be determined from the t-distribution tables in as much the same way as for simple regression explained above with the exception that in multiple regression, they would have a t-distribution with $n - k - 1$ degrees of freedom. The rule of thumb applicable in the case of large samples of datasets mentioned above, is also accepted by most researchers.

We shall now look at the output generated by Zoom Cars and interpret the given standard errors of coefficients.

⁹ University of Leicester, 2502 Financial Modelling, p.4.19

<u>Model 1</u>	Coefficient/Standard error (given)	t-statistic (see page 19)	(see $t_{\alpha/2, n-k-1}$ ¹⁰ 2.763(99%), 2.048(95%)
constant	-5.226/3.388	-1.5425	Not significant
price	-1.039/0.369	-2.8157	Significant
advertising	2.077/0.486	+4.2737	Significant
Favourable reviews	0.300/0.200	+1.5000	Not significant
M7	-1.185/0.054	-21.9444	Very Significant
M8	0.926/0.053	+17.4717	Very Significant
M12	-0.689/0.053	-13.000	Very Significant

The scope of this table is to help determine which variables are making significant contributions to the model. It transpires that the constant and the favourable reviews are not statistically significant while advertising seems to be very significant compared to price changes. The dummy variables also result to be highly significant in the model. Let us now analyse in the same way, model 2.

<u>Model 2</u>	Coefficient/Standard error (given)	t-statistic (see page 19)	(see $t_{\alpha/2, n-k-1}$ ¹¹ 2.763(99%), 2.048(95%)
constant	-4.612/2.785	-1.656	Not significant
price	-0.645/0.443	-1.4560	Not significant
advertising	1.038/0.677	+1.5332	Not significant
Favourable reviews	0.866/0.322	+2.6894	Significant
M7	-1.143/0.051	-22.4118	Very Significant
M8	0.979/0.056	+17.4821	Very Significant
M12			

Model 2 has resulted in a situation which does not lend much economic sense. One would expect price and advertising to contribute significantly to the model. However in this case, it has to be noted that these have been lagged by one month. The favourable reviews in the previous month motor press is, in this model a significant

¹⁰ t-distribution at 99% = 0.005, 95%=0.025, n=36 (given), k=7 (number of coefficients)

¹¹ t-distribution at 99% = 0.005, 95%=0.025, n=36 (given), k=7 (number of coefficients)

contributor, whereas in model 1 it was insignificant. The conflicting logic between these two models requires further analysis.

Goodness of fit of the regression line

The regression model shows that a change in y can be explained by a variation in the independent variable x and by the error term. The analyst will venture further into the realms of determining how much of the variation in y is caused by x and how much is caused by the error. In other words the analyst will find out how small the dispersion of data around the regression line is. The lesser the degree of dispersion of the data points in the scatter graph, the higher the degree of the true relationship demonstrated by the regression line. This is normally referred to as the goodness of fit.

A very common measure of the goodness of fit for regression models is the coefficient of determination, symbolised by R^2 . This coefficient will result in the proportion of the variation of the dependent variable explained by the independent variable and the result will vary between 0 and 1. An R^2 of 0 means that the equation explains none of the variation in the dependent variable and an R^2 of 1 means that the equation explains 100% of the variation in y . R^2 is calculated as follows:

$$R^2 = \frac{ESS}{TSS} = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2} \quad (\text{equation 11})$$

or

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2}$$

where ESS is the explained variation
 RSS is the unexplained variation or Residual Sum of Squares
 TSS is the Total Sum of Squares

The result of R^2 is often used in deciding whether or not to continue the analysis of a set of data. If the result of the test is a low percentage, say 10%, then it would be futile in trying to predict the future behaviour using the model.

The R^2 is often used to compare regression equations as long as the dependent variables in each equation being compared are identical and also that have the same number of explanatory variables.

In the case of Zoom Cars, RSS and TSS are computer generated and therefore the resultant R^2 will be as follows:

Model 1

$$1 - \frac{RSS}{TSS} = 1 - \frac{0.499}{4.302} = 88.4\%$$

Model 2

$$1 - \frac{RSS}{TSS} = 1 - \frac{1.465}{4.302} = 65.9\%$$

Part of the reason why the R^2 for these two models with similar variables has resulted in such different results, might be due to the fact that in model 1 there is one more explanatory variable than model 2 since in the latter case, M12 was not included. The additional explanatory variable will cause the coefficient of determination to increase and as a consequence, R^2 should be adjusted to take into account of the number of additional independent variables¹².

¹² Terry J Watsham & Keith Parramore, Quantitative Methods in Finance, p.203

The adjusted R^2 , symbolised as \bar{R}^2 is calculated as follows:

$$\bar{R}^2 = 1 - \frac{(1 - R^2)(n - 1)}{(n - k)} \quad (\text{equation 12})$$

This said about adjusted R^2 , if the inclusion of an additional variable in the model results in a decrease in \bar{R}^2 , the interpretation would be that the additional variable is not significant. However in effect, adding or deleting a variable is based on the theory behind the model not on the resultant movement in \bar{R}^2 .

Looking at the models in the Zoom Cars case and deriving the \bar{R}^2 for both will result as follows:

$$\text{Model 1} \quad \bar{R}^2 = 1 - \frac{(1 - R^2)(n - 1)}{(n - k)} = 1 - \frac{(1 - 0.884)(36 - 1)}{(36 - 7)} = 0.86$$

$$\text{Model 2} \quad \bar{R}^2 = 1 - \frac{(1 - R^2)(n - 1)}{(n - k)} = 1 - \frac{(1 - 0.659)(36 - 1)}{(36 - 6)} = 0.60$$

Given that the \bar{R}^2 is higher in the case of model 1 where the additional variable M12 is included, suggests that that variable is significant.

The coefficient of determination suggests that overall, the independent variables in model 1 lend a stronger explanation to the variation of the dependent variable. Model 1 is said to have more explanatory power than model 2. However, the coefficient still needs to be tested. Referring back to hypothesis testing explained briefly on page 18, the hypothesis to be tested now is as follows:

$$H_0 : R^2 = 0$$

$$H_0 : R^2 \neq 0$$

R^2 is itself a random variable. Therefore the test statistic is said to have an F distribution¹³. The F distribution is different from the other distributions in that it has two sets of degrees of freedom, one in the numerator of the test statistic and one in the denominator.

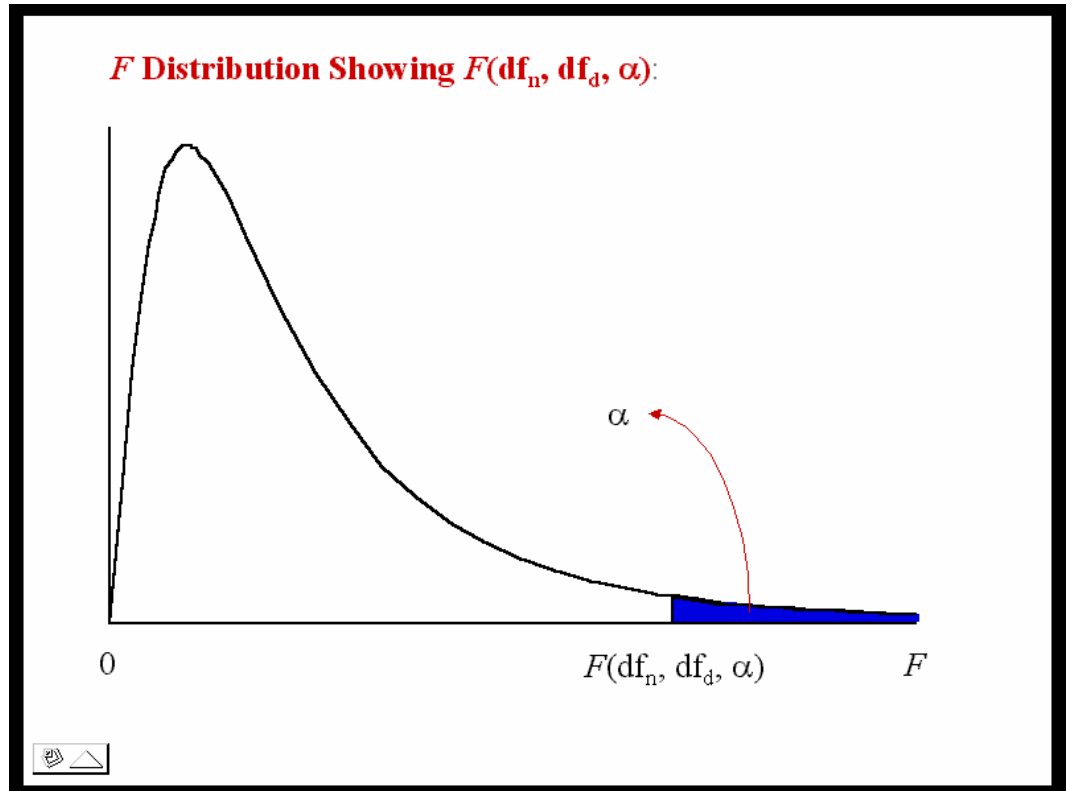


Diagram5 – F-distribution at df degrees of freedom and α level of significance¹⁴

The F-statistic is calculated as follows:

$$F = \frac{R^2}{\frac{(k-1)}{(1-R^2)} \frac{1}{(n-k)}}$$

¹³ Terry J Watsham & Keith Parramore, Quantitative Methods in Finance, p.200

¹⁴ Robert D Mason, Douglas A Lind, William G Marchal, Statistical Techniques in Business and Economics, 10th edition, Student CD-Rom by Irwin/McGraw-Hill

The degrees of freedom for the numerator are $(k-1)$ that is the number of coefficients less one, whereas that for the denominator is $(n-k)$ that is the number of observations less the number of coefficients. This test however only tells that there is correlation between y and x . It becomes particularly useful when testing the hypothesis for multiple regression with several variables, as in the case of Zoom Cars. The test statistic for the R^2 of the two models under interpretation is calculated as follows:

$$\text{Model 1} \quad F = \frac{\frac{R^2}{(k-1)}}{\frac{(1-R^2)}{(n-k)}} = \frac{\frac{0.884}{(7-1)}}{\frac{(1-0.884)}{(36-7)}} = 36.83$$

$$\text{Model 2} \quad F = \frac{\frac{R^2}{(k-1)}}{\frac{(1-R^2)}{(n-k)}} = \frac{\frac{0.659}{(6-1)}}{\frac{(1-0.659)}{(36-6)}} = 11.59$$

Referring to the F tables, we find that in the case of model 1, at 5% critical value for the degrees of freedom of 6, $(7-1)$, in the numerator and 29, $(36-7)$, in the denominator, the F value is 2.43. At 1% critical value the F value would be 3.5. Therefore, since the F statistic derived using the equation resulted in a value of 36.83, the null hypothesis, that is $H_0 : R^2 = 0$, is clearly rejected. This continues to support the results of the analysis that the equation of model 1 significantly explains variations in sales.

As for model 2 with 5 degrees of freedom in the numerator and 30 degrees of freedom in the denominator, the F tables would result in 2.53 at 5% critical value and 3.70 at 1% critical value. Again, the null hypothesis is rejected since the F statistic resulted in a value of 11.59 which is higher than the F values.

Conclusions

The a priori reasoning of the causality and elasticities of the variables, analysed on page 13, already gave model 1 an edge over model 2. The standard error of the equation analysed on page 16 also favoured model 1. The test on the goodness of fit of the two models on page 22 resulted in a rejection of the null hypothesis and that both models contribute to the explanation of variations in the dependent variable. The significance tests of the coefficients amplified on page 20 further supported the logic behind the reasoning of the causality referred to above. There seems to be a further link between this and the adjusted R^2 which hinted that the qualitative variable M12 has after all a significant effect on the model. All said, the analysis, testing and interpretation of the models presented by the analyst of Zoom Cars, clearly suggest that model 1 will be the better predictor of the dependent variable of the two models.

References

University of Leicester, '2502 Financial Modelling' , Edition 3, Learning Resources:
England

Terry J Watsham & Keith Parramore, 'Quantitative Methods in Finance', first edition,
Gray Publishing, Kent, England

LWT Stafford, 'Mathematics for Economists', first edition, Macdonalds and Evans Ltd,
England

William J Vincent, 'Statistics in Kinesiology', Human Kinetics, Leeds, England

Robert D Mason, Douglas A Lind, William G Marchal, 'Statistical Techniques in
Business and Economics', 10th edition, Student cd-rom by Irwin/McGraw-Hill

Mario F Triola, 'Elementary Statistics', 8th edition, StatSource