Chapter 7
Bivariate Correlation and Regression

An essential part of managing any organisation be it governmental, commercial, industrial or social, is planning for the future by generating adequate forecasts of factors that are central to that organisation's successful operation. Methods of forecasting fall into two groups; qualitative and quantitative. Among the former fall expert judgment and intuitive approaches. Such methods are particularly used by management when conditions in the past are unlikely to hold in the future. This chapter develop the quantitative regression approach to forecasting.

In all branches of scientific enquiry, no statistical tool has received the attention given to *regression analysis* during the last 30 years. A frequent misconception among non-statistical users of regression is that the technique's prime objective is to generate forecasts. Besides being a forecasting tool, regression analysis attempts to shed light on the mechanisms that relate variables. Knowledge of such mechanisms would, in some circumstances, permit a degree of control. For example, knowledge of how certain factors contribute towards the production of defective industrial items might assist in reducing the defective rate. Knowledge of the factors that drive changes in share prices would help in portfolio selection. Regression is a tool for forecasting and explanation. (Time series analysis, a technique that is discussed later, is solely a forecasting tool.) Of all regression models, the linear model is the most widely applied.

The next two sections review respectively bivariate (two-variable) correlation and linear regression. In terms of linear regression, Pearson's product-moment correlation coefficient is the most important measure of correlation. There are assumptions that underpin the regression method and which require attention before applying the method, even in the simple bivariate case. A fourth section illustrates the application of regression and correlation in IBM SPSS Statistics. Statistical and graphical plots are illustrated.

A. Aljandali, *Quantitative Analysis and IBM® SPSS® Statistics*,
Statistics and Econometrics for Finance, DOI 10.1007/978-3-319-45528-0_7

7.1 Bivariate Correlation

There are three types of bivariate correlation available in IBM SPSS Statistics; *Pearson's product moment correlation coefficient* (r), *Spearman's rank correlation coefficient* (ρ) and *Kendall's tau b*. These and all other correlation measures perform different tasks. Pearson's r is a measure of linearity and is thus the most important in relation to linear regression. In the bivariate case, if two variables X_i and Y_i ($i = 1, 2, \ldots, n$ where n is the number of observations) are not linearly related or correlated, there seems little point in proceeding with linear regression to generate forecasts. If the bivariate data points fall on a perfectly straight line with positive gradient, r would take the value 1; if they fall on such a line with negative gradient, r would be −1. If the Pearsonian coefficient is close to zero, it may well indicate an almost random scatter of the points. However, r being close to zero only indicates the absence of linearity. The variables at hand may exhibit a marked curved relationship. A U-shaped quadratic relationship would generate a value of r close to zero. Certainly in the bivariate case, it is wise to construct a scatterplot in IBM SPSS Statistics of the variables, to ensure that non-linear relationships are not ignored.

A common precursor to employing bivariate regression is to test whether or not the gathered sample data have been drawn from a population in which the Pearsonian correlation coefficient R is zero. As already stated, if the data have been drawn from such a population, then linear regression would seem a pointless exercise. The statistical test that:

H_0: sample is drawn from a population in which $R = 0$

is part of the IBM SPSS Statistics Correlations routine. The appropriate test statistic under H_0 is:

$$r\frac{\sqrt{n-2}}{1-r^2}$$

which is distributed as the t statistic with $n-2$ degrees of freedom; n is the number of sampled bivariate points. For example, suppose we were testing a one tailed alternative H_1: $R > 0$, then if $n = 47$ readings produced a sample correlation coefficient of $r = 0.34$, then the above equation would produce a test statistic with numerical value 2.4253. From statistical tables $P(t_{45} > 2.0141) = 0.05$, so we reject the hypothesis that $R = 0$. IBM SPSS Statistics prints exact significance levels for this test statistic.

The square of the Pearsonian coefficient (usually expressed as a percentage) is called the *coefficient of determination* $r^2\%$, so if $r = 0.8$ then $r^2\% = 64\,\%$. The coefficient of determination measures the amount of variation in the Y (the dependent variable) that is 'explained' by a linear relationship with the X (the independent variable). This parameter is a better indicator of the extent of linearity present than is r. For example, a sample correlation coefficient of $r = 0.7$ might seem relatively high. However, when viewed in the context of the coefficient of determination, we see that less than half of the variation in the Y is explained by a linear relationship with the X. The 51 % of variation in the Y variable that is unexplained could be due to errors in measurement and to variables that have not been included in the study. The latter variables are called *extraneous factors* and would be of consequence in the above example.

Spearman's rank correlation by its very title is a bivariate measure of association when the data comprise ranks, for example, two consumers might rank five competing brands from 1 ("most preferred") to 5 ("least preferred"). Spearman's ρ measures the degree of similarity between the consumers' rank scores. In fact, it may be shown that the numerical value of Spearman's ρ is the same as the value that would be obtained if the ranked scores were inserted into the formula for Pearson's r. Spearman's ρ is thus a measure of the degree of linearity between the ranked scores. If the two consumers agreed exactly in their rankings of the brands, then these ranks would be perfectly linearly related and $r = 1$. If the ranks were the exact opposite of each other, then $r = -1$.

Just as in the cases of Spearman's coefficient, Kendall's tau b is a measure of association that applies to ordinal data and may take values between ±1 inclusive. Essentially, the latter statistic looks at all possible pairs of cases and examines whether or not the ranks are in the same order.

7.2 Linear Least Squares Regression for Bivariate Data

If the coefficient of determination indicates that there is adequate linearity present in the gathered bivariate data set (e.g. $r^2 > 80\,\%$), then we may proceed to fit a line through the data points and use this line for forecasting purposes. This line is called *the linear least squares regression line*, denoted in the bivariate case by $\hat{Y} = a + bX$, where $\hat{Y}$ represents the forecasted or predicted values. The method for determining the optimal value for the parameters a and b involves minimizing the sum of squares of the errors (or *residuals*) about the regression line i.e.:

$$\min \sum \left(Y - \hat{Y} \right)^2$$

where the expression $(Y - \hat{Y})$ is the difference between the observed and predicted values of Y, namely the *error* or *residual*. We are unable to minimise the sum of the residuals themselves, because this sum is always zero. Gathered data are most often a sample from a particular population. The population regression line is denoted by $Y = \alpha + \beta X$. Given the importance of gradients in fields such as Economics (in that gradients reflect rates of change; for example), it is often desirable to make inferences about β based on the value of b derived from the sample data. In particular, it is most frequently hoped that we have sufficient sample evidence to reject:

$$H_0 : \beta = 0.$$

For example, suppose it is believed that revenue (Y) in sales regions depends on the number of television advertisements (X) shown in those regions over a particular period. If we failed to reject the above null hypothesis, this would infer that the population gradient is zero, i.e. for every advertisement shown, there is no increase in sales revenue.

The above null hypothesis is assessed via the test statistic:

$$(\mathrm{b}-\beta)/(\mathrm{SE\,of\,b}),$$

where SE represents *standard error*. Recalling that statistical tests are conducted under the assumption that the null hypothesis is true, the value of β in the above would be zero. However, if one wished to test H_0: $\beta = 4$ (e.g. an accountant may wish the rate of return on a project of at least four units per financial unit invested), then β would take a value of four in the formula for the test statistic. The above test statistic is distributed as the t statistic with $n-2$ degrees of freedom. The method of calculating the standard error of b is tedious and requires some statistical expertise, so interested readers are referred to standard statistical texts.

7.3 Assumptions Underlying Linear Least Squares Regression

There are a series of statistical assumptions that underpin the least squares regression model and which will influence our selections from the IBM SPSS Statistics dialogue boxes illustrated in the next subsection. Most of the assumptions refer to the residuals or errors from the regression analysis. Remember that a residual is simply the difference between the observed value of the dependent variable at a particular value of X and the predicted value of the dependent variable at that value of X.

The assumptions that underlie regression are:

- the relationship between the Y and X should be adequately approximated by a straight line,
- the residuals should have zero mean,
- the residuals should have constant variance (called the *homoscedasticity* assumption),
- the residuals should not be correlated and
- the residuals should be normally distributed.

A scatter diagram will help to indicate the adequacy or otherwise of using a linear regression approach. In the bivariate case, the coefficient of determination gives a less subjective approach. The residuals will have a zero mean due to the mechanics whereby the least squares line is computed.

If the homoscedasticity assumption is violated, the constants a and b (called *regression coefficients*) in the regression equation are still unbiased estimates of their population equivalents α and β, but they are no longer minimum variance estimators. Put more simply, if the variance or spread of the errors is, for example, increasing as we move along the regression line, then why are we fitting a line to the data? Certainly, forecasts into the future are likely to possess high error in this situation. A simple way of assessing the constant residual variance assumption is graphical. A plot of the residuals against the predicted values $\hat{Y}$ (i.e. the regression line) should show no evidence of varying spread.

Correlation between the residuals is common when we have economic data recorded over time. This phenomenon is called *temporal autocorrelation* and again may be assessed graphically. A plot of the residuals over time should evidence a random pattern i.e. the residuals are independent of time. Again, the regression coefficients are still unbiased but they are no longer minimum variance estimates in the face of autocorrelation. Confidence intervals for α and β tend to be narrower than they should be. (There is a formal test for autocorrelation called the *Durbin–Watson test* and which is available in IBM SPSS Statistics.)

It should be noted that small departures from normality of the residuals does not affect the linear regression model greatly. However, gross violations of this assumption can seriously distort the test statistic for the population gradient β and associated confidence intervals. The normality of the residuals may be assessed via a normal probability. Should there be sufficient residuals, then a histogram will give a graphical assessment of this assumption. It is also possible to save the residuals in the existing IBM SPSS Statistics data file and subject them to the Shapiro–Wilks available in the Explore procedure (Sect. 3.2).

This subsection has indicated that there are various plots that may assist in the assessment of the regression assumptions and naturally, these plots are available in IBM SPSS Statistics. By default, such plots *standardize* the variables in question, such as the residuals. This is to remove the problem of the unit of measurement. For example, if we had a residual or error of $50,000,000 in a problem, this may seem immense. However, if the problem involved forecasting the gross national product of the U.S.A., then the residual is far from serious! Standardization of residuals overcomes the problem of units. Standardization leads the residuals to have a mean of zero and variance of one.

7.4 Bivariate Correlation and Regression in IBM SPSS Statistics

The file EARNINGS.SAV contains the returns on assets for 30 American firms in terms of several potential explanatory variables. These variables along with their IBM SPSS Statistics names are:

- CAPGWTH—capital spending growth,
- INDRET—industry return on assets,
- INDSALES—industry sales growth,

- INVTURN—inventory turnover,
- OPINCGTH—operating income growth,
- OPINCRAT—the ratio of operating income to total assets and
- SALESGTH—sales growth.

The first stage in the presented analysis of this section is to see which of these seven variables has the strongest linear relationship with the variable to be explained, COASSETS. This will be achieved by running the IBM SPSS Statistics bivariate correlations procedure.

To access the *Bivariate Correlations dialogue box* of Fig. 7.1, from the Data Editor click:

Analyze
Correlate
Bivariate…

and you will see the availability of the three bivariate correlation measures discussed in Sect. 7.1, with Pearson being the default. A two-tailed test of significance (the default) has been selected in this case. This option should be used when the user is not in a position to know in advance the direction (positive or negative correlation) of the relationship.

The dialogue box of Fig. 7.1 in fact generates the value of Pearson's r for all pairs of the eight variables in the box titled 'Variables'. Figure 7.2 presents the statistical output. The variable COASSETS ('FIRM'S RETURN ON ASSETS') is regarded

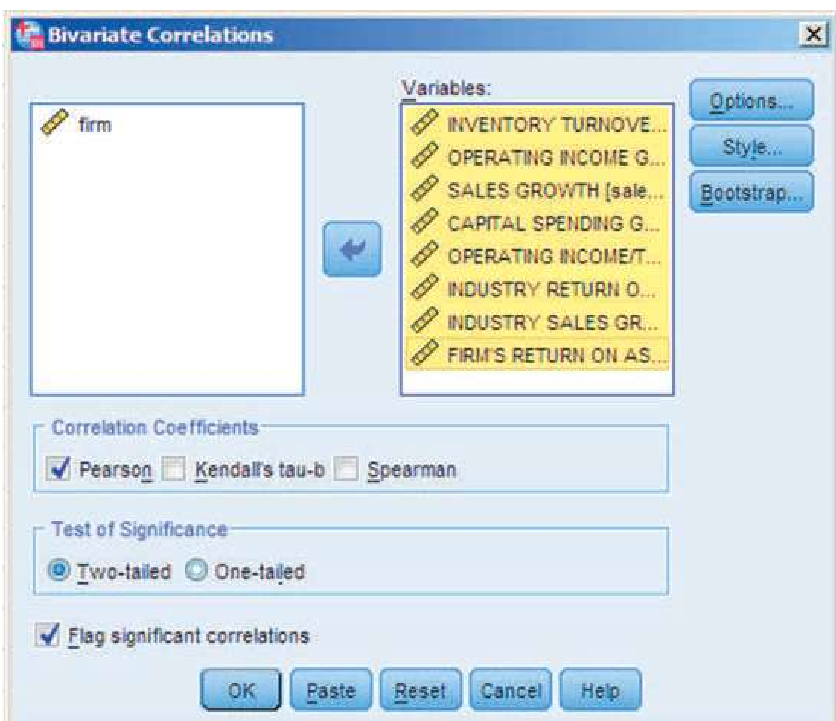


Fig. 7.1 The Bivariate Correlations dialogue box

Correlations

		INVENTORY TURNOVER	OPERATING INCOME GROWTH	SALES GROWTH	CAPITAL SPENDING GROWTH	OPERATING INCOME/TOTAL ASSETS	INDUSTRY RETURN ON ASSETS	INDUSTRY SALES GROWTH	FIRM'S RETURN ON ASSETS
INVENTORY TURNOVER	Pearson Correlation	1	-.213	-.484**	-.270	-.389*	-.608**	-.752**	-.515**
	Sig. (2-tailed)		.257	.007	.149	.034	.000	.000	.004
	N	30	30	30	30	30	30	30	30
OPERATING INCOME GROWTH	Pearson Correlation	-.213	1	.443*	.138	-.106	.091	.310	.047
	Sig. (2-tailed)	.257		.014	.468	.578	.634	.096	.804
	N	30	30	30	30	30	30	30	30
SALES GROWTH	Pearson Correlation	-.484**	.443*	1	.334	.351	.289	.453*	.250
	Sig. (2-tailed)	.007	.014		.072	.057	.122	.012	.183
	N	30	30	30	30	30	30	30	30
CAPITAL SPENDING GROWTH	Pearson Correlation	-.270	.138	.334	1	.282	.154	.189	.311
	Sig. (2-tailed)	.149	.468	.072		.132	.416	.318	.094
	N	30	30	30	30	30	30	30	30
OPERATING INCOME/TOTAL ASSETS	Pearson Correlation	-.389*	-.106	.351	.282	1	.347	.431*	.763**
	Sig. (2-tailed)	.034	.578	.057	.132		.060	.017	.000
	N	30	30	30	30	30	30	30	30
INDUSTRY RETURN ON ASSETS	Pearson Correlation	-.608**	.091	.289	.154	.347	1	.793**	.380*
	Sig. (2-tailed)	.000	.634	.122	.416	.060		.000	.039
	N	30	30	30	30	30	30	30	30
INDUSTRY SALES GROWTH	Pearson Correlation	-.752**	.310	.453*	.189	.431*	.793**	1	.410*
	Sig. (2-tailed)	.000	.096	.012	.318	.017	.000		.025
	N	30	30	30	30	30	30	30	30
FIRM'S RETURN ON ASSETS	Pearson Correlation	-.515**	.047	.250	.311	.763**	.380*	.410*	1
	Sig. (2-tailed)	.004	.804	.183	.094	.000	.039	.025	
	N	30	30	30	30	30	30	30	30

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Fig. 7.2 Output from running bivariate correlation as the dependent variable. Figure 7.2 shows that COASSETS is significantly correlated with the INVTURN. The value of the Pearsonian correlation between COASSETS and INVTURN is −0.515 for the 30 sample observations. The significance associated with the null hypothesis:

H_0: the population correlation coefficient, R, between these two variables is zero is $p=0.004$. The null hypothesis is, therefore, rejected and the correlation is significantly different from zero. The variable COASSETS is also significantly correlated with INDSALES, as indicated by the levels of significance being less than 0.025 for these two tailed tests. All levels of significance are based on computation of the test statistic and the t distribution with $n-2$ degrees of freedom.

The variable OPINCRAT, in fact, has the most significant correlation with COASSETS and will, therefore, be used to illustrate bivariate linear regression. The correlation exercise, therefore, suggests that the variable OPINCRAT is the most important determinant of the values of the variable COASSETS.

From the Data Editor, click:

Analyze
  Regression
    Linear…

to obtain the *Linear Regression dialogue box* of Fig. 7.3. The 'Dependent' variable in this example is COASSETS which is entered into the associated box in the usual manner. The 'Independent' variable is OPINCRAT. There are several methods for conducting regression analysis, most of which are pertinent if the researcher is

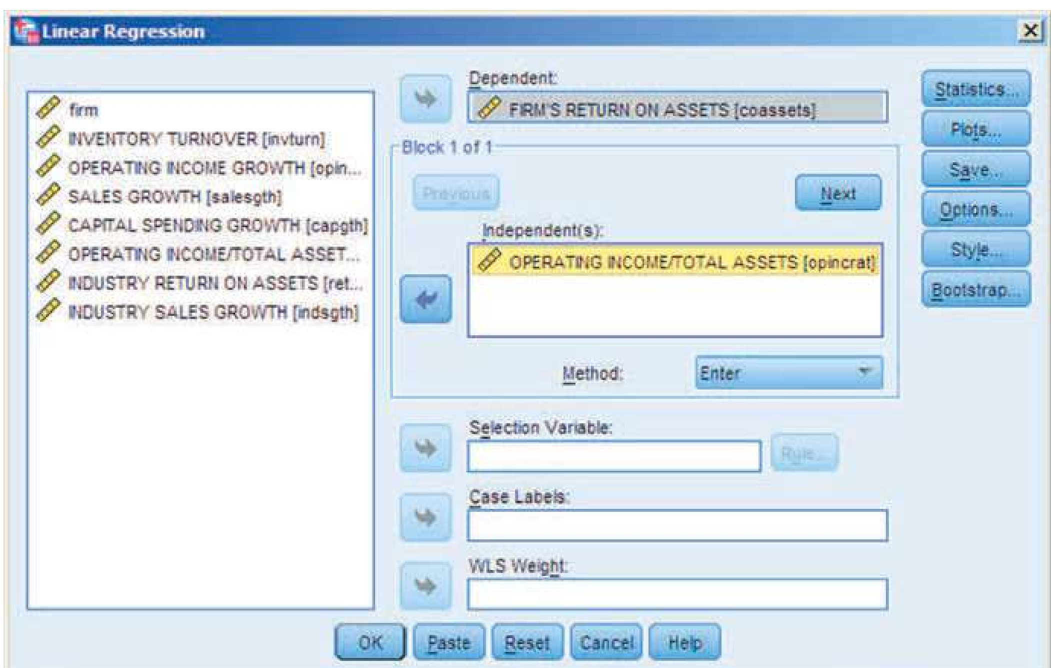


Fig. 7.3 The Linear Regression dialogue box

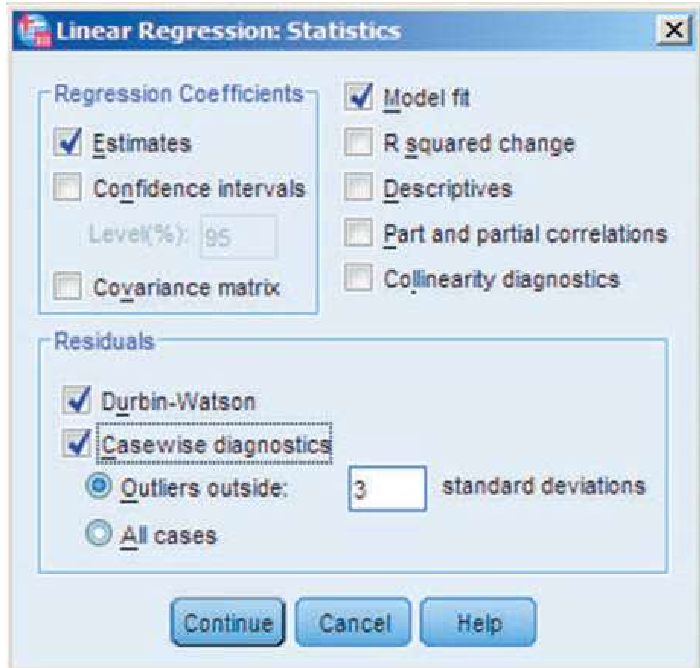


Fig. 7.4 The Linear Regression: Statistics dialogue box

pursuing a multivariate analysis. Suffice it to say at present that the default procedure in the 'Method' box of Fig. 7.3 is called the "Enter Method" and will be chosen here. Generally, this method enters all of the independent variables in one step. Here, of course, we only have one independent variable.

Click the Statistics… button at the top right hand corner of the *Linear Regression dialogue box* to obtain the *Linear Regression: Statistics dialogue box* of Fig. 7.4. By default, estimates of the regression coefficients are produced. Confidence intervals for these coefficients are optional as are various descriptive statistics. Note that the Durbin–Watson test for autocorrelation is selected from this dialogue box if desired, but our COASSETS data are not temporal, so this test is irrelevant. *Casewise diagnostics* (i.e. firm by firm) of standardized residuals for all cases has been chosen from this dialogue box.

Note that under the heading 'residuals', this dialogue box accommodates the detection of 'outliers'. Loosely speaking, *outliers* are points that are far distant from the regression line i.e. they have large positive or negative residuals. They could represent data input errors. They could also be points of special interest that merit further study or separate analysis. Recall that IBM SPSS Statistics standardizes the residuals (mean of zero and variance of one). By default, points more than three standard deviations either side of the regression line are regarded as outliers in IBM SPSS Statistics. The user may change this limit in the dialogue box of Fig. 7.4. Click the Continue button to return to the *Linear Regression dialogue box* of Fig. 7.3.

At the bottom of the *Linear Regression dialogue box* is the Plots… button that gives rise to the *Linear Regression: Plots dialogue box* of Fig. 7.5, which permits graphical evaluations of the assumptions underlying the regression method and

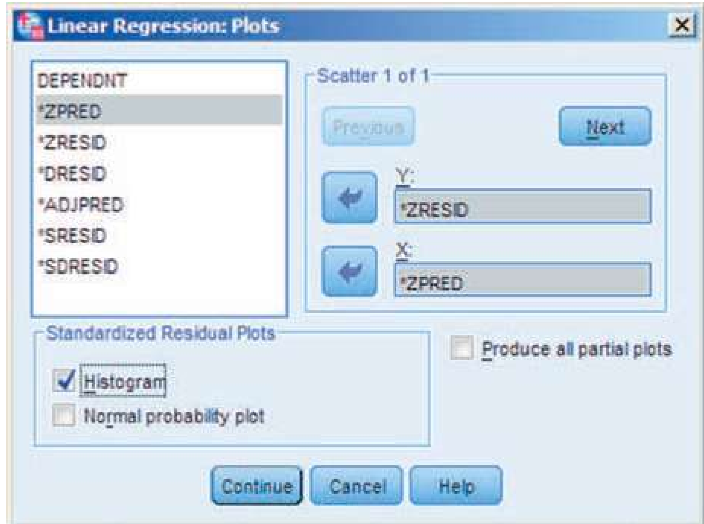


Fig. 7.5 The Linear Regression: Plots dialogue box

which were discussed in Sect. 7.2. A plot of the (standardized) residuals against the (standardized) predicted values allows the research to judge if homoscedasticity is present. In IBM SPSS Statistics, the standardized residuals are denoted by *ZRESID and the standardized predicted values by *ZPRED. These are respectively clicked into the boxes labelled Y and X, as shown, via the arrow buttons. In Fig. 7.5, a histogram has also been selected to assess the normality assumption pertaining to the residuals.

Click the Continue button to return to the *Linear Regression dialogue box* of Fig. 7.3. Again at the top right of this dialogue box is the Save… button which accesses the *Linear Regression: Save dialogue box* of Fig. 7.6. Many of the options here require advanced knowledge of regression techniques. However, the user may wish to save 'Unstandardized' and 'Standardized' predicted and residual values for further study or graphical analysis. The appropriate boxes are simply clicked and a cross appears in each upon selection. Click the Continue button to return to the *Linear Regression dialogue box* and then the OK button to perform the regression analysis. Figure 7.7 presents part of the results of this regression analysis in the IBM SPSS Statistics Viewer.

The value of the coefficient of determination is $r^2\,\% = 58.2\,\%$, so nearly 40 % of the variation in company COASSETS remains unexplained after the introduction of the operating income/total assets ratio. Clearly, some of the other variables that are significantly correlated with these firms' returns should be introduced into our analysis, creating a multivariate regression problem. The value of the Pearsonian correlation between the variables COASSETS and OPINCRAT is 0.763 ($p = 0.000$).

The equation of least squares linear regression is:

$$\text{COASSETS} = 11.697 + .639 * (\text{OPINCRAT}),$$

but this bivariate equation of regression would, in all probability, be inadequate for forecasting purposes in that r^2 is not sufficiently high. The above figure permits

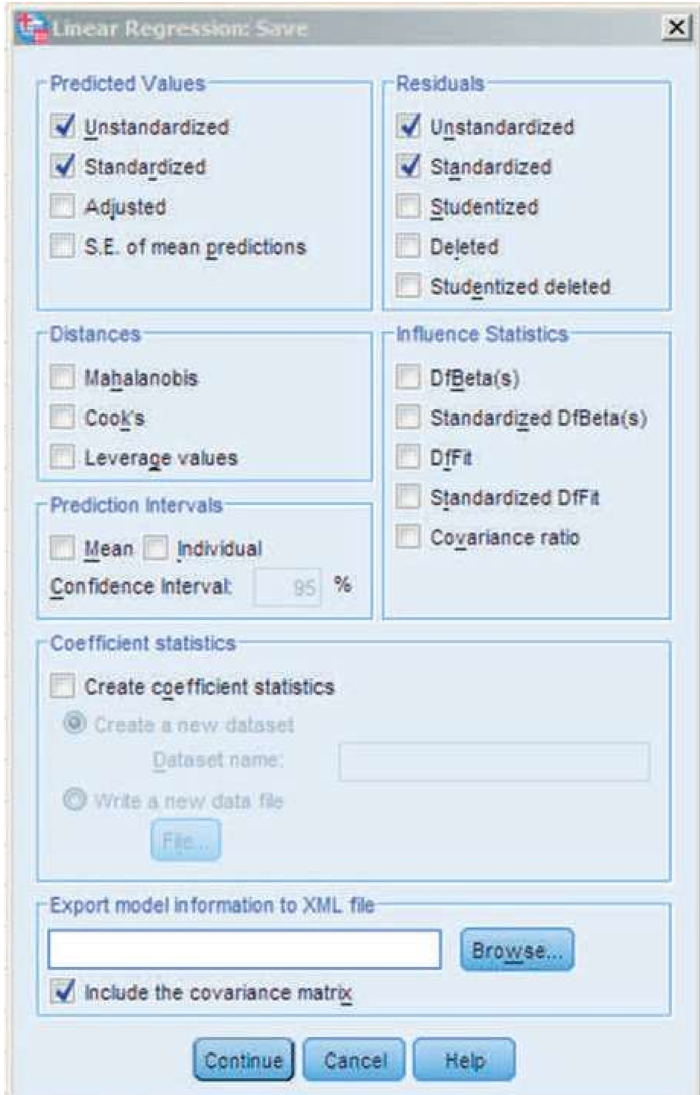


Fig. 7.6 The Linear Regression: Save dialogue box

study of the hypothesis that the population regression line has a zero gradient (i.e. H_0: $\beta=0$). From Fig. 7.7, we, therefore, derive a test statistic of:

$$\mathrm{b}/(\mathrm{SE\ of\ b})=0.639/0.102=6.264(\text{remember } \beta=0 \text{ under the null hypothesis}),$$

which is distributed as a t statistic with $n-2=28$ degrees of freedom. This test statistic is part of the output of Fig. 7.7 and has a significance level of $p=0.000$ to three decimal places. We thus reject the null hypothesis and conclude that the population gradient is non-zero. Our best estimate of β is simply the sample value of 0.639. This means that on average, a one unit increase in the operating income to total assets ratio generates a 0.639 increase in company returns. A 95 % confidence interval for the population gradient is in fact given by:

$$P(0.429<\beta<0.848)=0.95$$

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.763[a]	.582	.567	4.3907	2.196

a. Predictors: (Constant), OPERATING INCOME/TOTAL ASSETS

b. Dependent Variable: FIRM'S RETURN ON ASSETS

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	11.697	3.700		3.161	.004
	OPERATING INCOME/TOTAL ASSETS	.639	.102	.763	6.243	.000

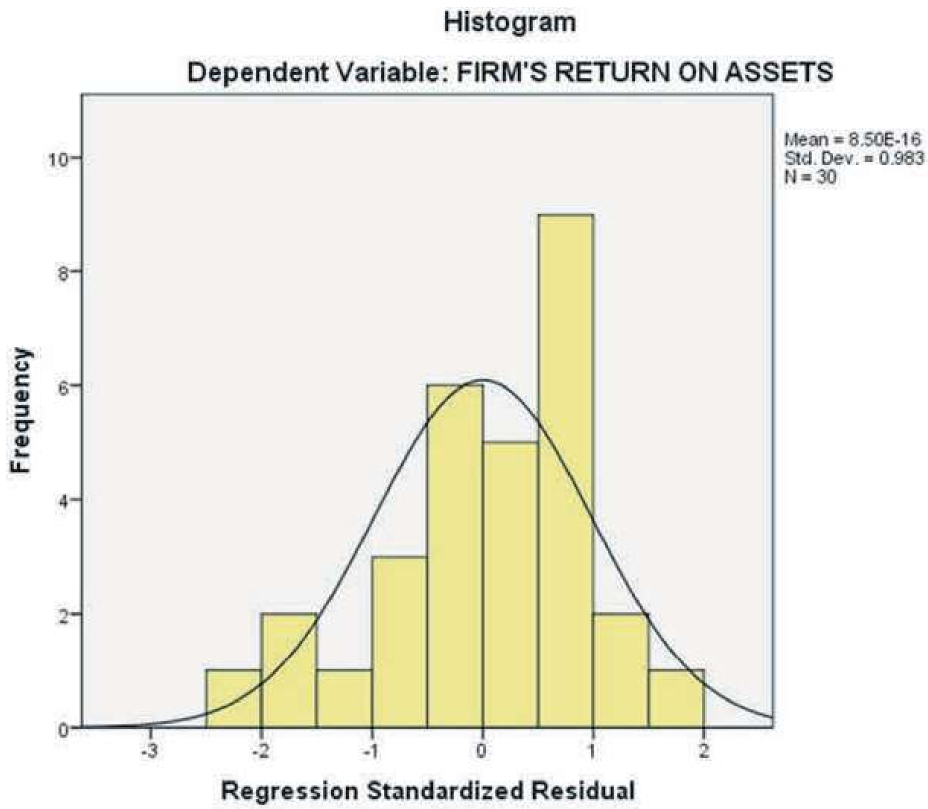
a. Dependent Variable: FIRM'S RETURN ON ASSETS

Casewise Diagnostics[a]

Case Number	Std. Residual	FIRM'S RETURN ON ASSETS	Predicted Value	Residual
1	-.387	31.9	33.601	-1.7008
2	.741	42.6	39.348	3.2519
3	1.348	42.2	36.283	5.9172
4	.624	35.0	32.260	2.7403
5	-1.441	24.4	30.727	-6.3271
6	.156	36.2	35.517	.6835
7	.995	36.5	32.132	4.3680
8	.530	33.5	31.174	2.3259
9	.307	34.5	33.154	1.3462
10	-.388	30.3	32.004	-1.7043
11	-2.063	38.4	47.458	-9.0582
12	.292	31.5	30.216	1.2838
13	.736	30.7	27.470	3.2297
14	.997	44.3	39.923	4.3772
15	-.197	32.8	33.665	-.8646
16	-1.934	17.7	26.193	-8.4931
17	-.738	31.0	34.239	-3.2394
18	.117	35.2	34.686	.5136
19	.626	38.2	35.453	2.7473
20	-.402	28.9	30.663	-1.7632
21	-.313	32.1	33.473	-1.3730
22	1.323	44.9	39.093	5.8073
23	-.924	38.1	42.158	-4.0579
24	.228	24.0	23.000	.9999
25	-.977	33.4	37.688	-4.2878
26	1.654	46.8	39.540	7.2603
27	-.332	27.1	28.556	-1.4559
28	.668	34.3	31.366	2.9343
29	-1.835	28.1	36.155	-8.0551
30	.591	42.9	40.306	2.5940

a. Dependent Variable: FIRM'S RETURN ON ASSETS

Fig. 7.7 Part of the output from running bivariate regression



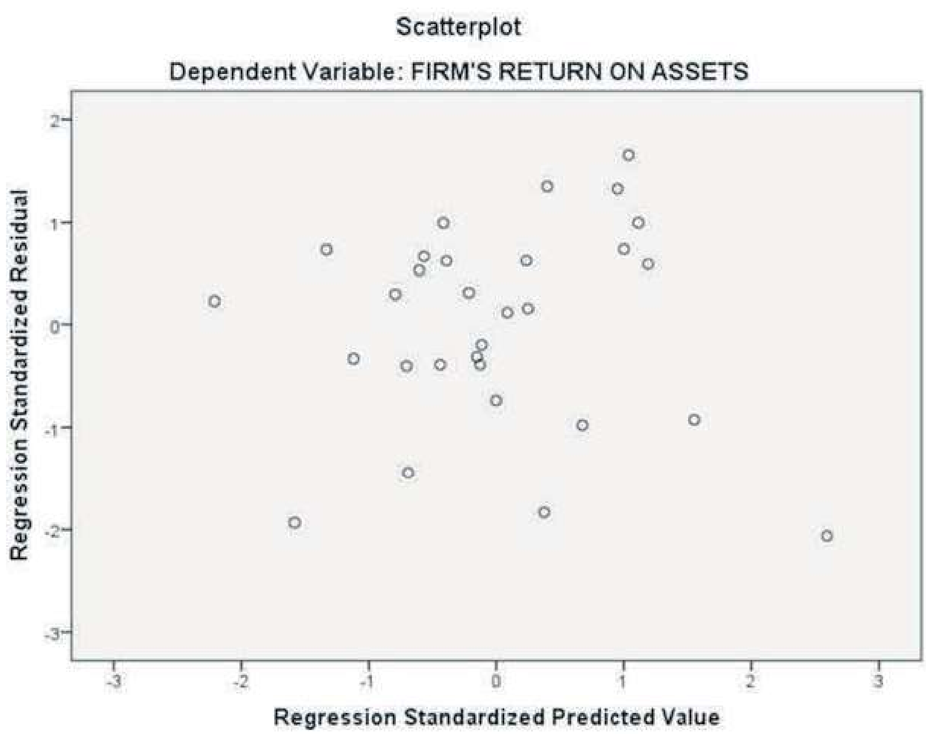


Fig. 7.7 (continued)

and note that a value of $\beta = 0$ is not contained in this interval, as is expected after conducting the hypothesis test on the population gradient. The *beta coefficient* (=0.763) reported in Fig. 7.7 is the coefficient of the independent variable when all variables are expressed in standardized (Z score) form. In a multivariate problem, it would be wrong to compare all the regression coefficients as indicators of the relative importance of each independent variable, since the size of a regression coefficient depends on its unit of measurement. Beta coefficients assist in the comparison process by means of standardization.

(It also possible to test H_0: $\alpha = 0$ i.e. the population intercept is zero via the test statistic:

$$\text{a} / \left(\text{Standard Error of a}\right),$$

which is also distributed as a t statistic. Here $t = 3.161$, $p = 0.004$, reject the null hypothesis. Consideration of the intercept, however, namely when the value of COASSETS when OPINCRAT equals zero has little relevance in this particular context.)

Included in Fig. 7.7 is information pertaining to the predicted and residual values obtained. This Figure indicates a maximum absolute standardized residual of 2.362 associated with company number 14. If we consider three standard deviations away from the regression line as the characteristic of an outlier, our study has thrown up no outliers. However, some researchers consider cases that are over two standard deviations away from the regression line as outliers, in which instance companies numbered 14 and 20 would be so considered.

The histogram (*on page 131*) suggests that the residuals show some departures from normality. The second diagram of Fig. 7.7 might suggest that the spread of the residuals is increasing as we move along the regression line. There may well be a problem as regards the homoscedasticity assumption.

An outward-opening funnel pattern on such a plot of is symptomatic of violation of the constant variance assumption. (The usual method for dealing with violation of this assumption is to use a method called *weighted least squares* and which is available in IBM SPSS Statistics.)

To summarize, the assumptions underlying regression and that relate to the bivariate case seem violated in terms of the requirements of homoscedasticity and normality of residuals. We have no outliers that would represent firms exhibiting non-typical behaviour in terms of the variables examined. However, the coefficient of determination should be higher and company returns on assets are inadequately explained by simply the operating income to total assets ratio. Forecasting the returns on assets of other companies using our derived regression line would probably be prone to unacceptable error. We, therefore, need to treat the analysis in a multivariate manner and introduce other, salient independent variable(s).