CASE STUDY 3

Correlation and Regression*

Problem

An organisation would like to build a regression model consisting of four independent variables to predict the compensation (dependent variable) of its employees. Past data has been collected for 15 different employees, and four independent variables. Build a regression model and recommend its proper usage.

The data is as follows:

Dependent variables

Y = compensation in rupees.

Independent variables

1. Experience (in years)
2. Education (in years, after 10th standard)
3. Number of employees supervised
4. Number of projects handled

The dataset consisting 15 observations, is given in Table 1.1

Correlation

The correlation table is shown in Table 1.2. The values in the correlation table are standardised and range from zero to one, positive and negative. Looking at the last column, we can say that all the variables are highly correlated to compensation ranging from 0.73 to 0.95 except experience whose correlation is 0.6049. This means that the independent variables have been chosen in a fairly good manner. This correlation shown in Table 1.2 is a one-to-one correlation of each variable with the other, so we still have to do multiple regression with an independent variable showing low correlation with a dependent variable because in the presence of other variables this independent variable may become a good predictor of the dependent variable.

The other point to be noted in the correlation table is whether independent variables are highly correlated with each other. In certain cases, they are highly correlated. This may indicate that they are not independent of each other and we may be able to use only 2 or 3 of them to predict the dependent variables.

Regression

We will first run the regression model of the following form by entering all the four X variables in the model.

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$$

and determine the values of the above.

Output values are:

$$a = 456.84$$
$$b_1 = 110$$
$$b_2 = -93.43$$
$$b_3 = 44.21$$
$$b_4 = 77.51$$

These values can be substituted in the above equation and the equation can be written as follows

Sales = 456.84 + 110 (education) – 93.43 (experience) + 44.21 (no. of supervisors) + 77.51 (no. of projects)

Now we need to look the statistical significance of the model and the R^2 value. These are available from Table 1.3, the analysis of variance table and Table 1.4, the regression output. From Table 1.3, the last column indicates the p-level to be 0.0001. This indicates that the model is statistically significant at a confidence level of (1 – 0.0001) 100 or .999 × 100 which is equal to 99.9.

The R^2 value is 0.88764 from the top of Table 1.4. From the same figure we can also note that t-test for the significance of independent variables indicate that at a significance level of 95% only experience, education, and number of projects are statistically significant in the model. The number of employees supervised is not significant.

There is one negative coefficient, that of experience, which can be interpreted to mean that if we increase the employee's experience, the compensation will decrease (according to – 93.43 coefficient of the variable).

We have another independent variable, that is the no. of people supervised, that shows a 't' value as .4459, that is, this value is statistically not significant.

Therefore education, experience, and number of projects are significant, and should be used for the interpretation. Therefore one should look at these to determine the compensation.

We can use the forward stepwise regression method or the backward stepwise regression method to try and eliminate the insignificant variable from the full regression model containing all the four independent variables.

Forward Stepwise Regression

Table 1.5 shows the result of forward stepwise regression, which ends up with only three out of four independent variables of the regression model. The three variables are projects (no. of projects handled), experience, and education. We notice that these variables are statistically significant at 95% confidence level.

F-Test of the model is also highly significant and R^2 value is 0.88057. We need not take the no. of people supervised in case we decide to use this model. If we decide to use this model it would be written as follows.

$$Y = 462.79 + 114.344 \text{ (education)} - 82.08 \text{ (experience)} + 89.18 \text{ (projects)}.$$

Backward Stepwise Regression

Table 1.6 shows the output for the backward stepwise regression. The results show that only education, experience, and projects remain in the model. R^2 value is 0.88057. The F-test for the model is highly significant and the independent variables are statistically significant at 95% confidence level. P-levels = 0.0323, 0.0456, 0.0031.

If we were to decide to use this model for prediction, we only require the data to be collected on the above 3 independent variables. We could form the equation as

$$Y = 462.79 + 114.344 \text{ (education)} - 82.08 \text{ (experience)} + 89.18 \text{ (projects)}.$$

Input Data

TABLE 1.1

	Compensa	Experien	Educatio	Nosuper	Projects
1	1500.00	2.00	5.00	4.00	10.00
2	1650.00	3.00	6.00	5.00	10.00
3	1750.00	3.00	3.00	5.00	12.00
4	1400.00	2.00	3.00	3.00	9.00
5	2000.00	4.00	4.00	6.00	15.00
6	2200.00	5.00	6.00	6.00	14.00
7	2100.00	1.00	5.00	4.00	12.00
8	2750.00	5.00	8.00	7.00	15.00
9	2900.00	8.00	9.00	8.00	25.00
10	1100.00	3.00	3.00	2.00	7.00
11	1000.00	4.00	2.00	1.00	5.00
12	1350.00	6.00	4.00	4.00	12.00
13	1550.00	4.00	6.00	4.00	11.00
14	1375.00	8.00	4.00	8.00	13.00
15	1400.00	4.00	3.00	5.00	10.00