IIM - Rohtak

Correlation and Regression: A Teaching Note

02/2018

This teaching note is prepared by Dr. S. K. Pandey, Assistant Professor of Marketing at Indian Institute of Management Rohtak. It is intended to be used as the basis for class discussion.

Correlation and Regression: A Technical Note

If there are two variables which are metric (interval scales like Likert or ratio scales such as height, weight, length, sales, etc.), correlation answers whether the two have a relationship. Recall the famous dialogue from Sholay when Gabbar is asking Basanti to dance and Veeru interjects. Gabbar says, “Bada Yaarana hai”. In marketing analytics, yaarana can be replaced by “correlation” in metric variables.

The first question is whether there is any relationship. If no, the two entities do not effect each other (indifferent). If yes, then after a relationship has been confirmed, the next thing to check is the direction of the relationship. Who says hate is not a relationship, it is also a relationship in a negative direction. The two variables may be positively related, meaning when one increases the other also increases. Eg. Height and weight. In marketing, the advertising budget and brand recall etc.

The two variables may be negatively related. The most recallable concept is the law of demand. When price increases quantity demanded decreases, of course, *ceteris paribus*.

Karl Pearson Correlation coefficient is the measure of relationship between two metric variables like height and weight, sales promotion and sales etc. The value of this coefficient varies between -1 and 1. A zero value indicates no *LINEAR* relationship meaning although the two variables don’t have any linear relationship they may still have a nonlinear relationship like quadratic, exponential etc.

After there is a relationship and we know whether it’s positive or negative, we need to predict one variable if we know the other. The square of R, R^2 is called, coefficient of determination. This R^2 explains how much of the variation in one variable can be determined by the other variable.

There is another coefficient called, Spearman’s Rho, used to measure relationship between two ordinal (rank) variables. Let’s take one example: Suppose MMI course is taught by two professors, Professor A and Professor B to the same students. Professor A teaches pre midterm and Professor B teaches post midterm. The performance of students is likely to be correlated in the two editions.

Our hypothesis is no linear correlation i.e. R=0

The data is metric i.e. marks (ratio scale) and we can find Karl Pearson Correlation Coefficient.

In SPSS go to Analyse/Correlate/Bivariate and then put the two variables in the test box and click OK.

Firstly we have to see whether we can say with more than 95% confidence that there is a correlation between the two variables. The p value if less than .05 then correlation is significant at 95% and if the p value is less than .01, then the p value is significant at 99%. The good thing is that the significant correlations are flagged. If the correlation is significant at 95%, one star will appear next to the Pearson Correlation Coefficient and if it is significant at 99% then two stars will appear.

If none of the stars appear then the null hypothesis of NO Correlation has to be accepted and no further inference can be drawn no matter what the R value is. If the correlation is significant (shown by stars), then the next thing is to look at the sign of the correlation coefficient. A positive value means a direct relationship and a negative sign means an inverse relationship. The third thing to look now is the R^2 value. The R^2 will always be positive and as mentioned earlier, it tells the amount of variation in one variable explained by the other. This is an indication of the strength of relationship.

However, if the data is ordinal, meaning instead of marks the two faculties gave ranks to the students. Then we can calculate Spearman's Rho and the interpretation is exactly like Pearson.

The data entered will be like

Student	Pre Midterm Marks	Post Midterm Marks	Pre Midterm Rank	Post Midterm Rank
1	25	36	17	8
2	84	72	1	2
3	33	15	10	19
4	79	63	4	5
5	58	58	7	6

We will do Karl Pearson Correlation in the first two variables and Spearman's Rho in the last two variables.

Regression

Recall that once we know that a relationship exists, we would like to predict the behaviour of one variable once we know the other. I will tell you a real incident. I was in one of the IP affiliated colleges and a lady Accounts lecturer was taking attendance. I happen to be present there. She happened to call a girl's name and she was absent. The lady lecturer stopped taking attendance, looked at the class, smiled and said that then a boy, named X, would also be absent and everyone started laughing. But she was absolutely following regression after knowing correlation. Once there is a relationship between the two variables, the behaviour of one variable can be predicted if we know the other variable. So simply saying regression is the average relationship between the two variables. Regression is meaningful only when the two variables are related (meaning correlation is significant) hence correlation is the necessary condition before regression. It is this reason that correlation and regression are always pronounced together. In marketing, if we found that advertising budget affects sales, the manager would like to know how much increase in advertising will lead to how much sales. Another example could be negative relationship in price and demand. The objective of a manager is not to reduce or increase price or not to reduce or increase demand, the objective often is to maximize profit. So regression can give the answer that if there is a relationship between price and quantity demanded, then how much change in price will lead to how much change in demand and vice versa.

So, the most important objective in regression is to predict the value of the other when we know one. First we will take a case of only two variables called bivariate (bi=two variables), one is a dependent variable and the other independent variable.

The objective is to predict the dependent variable with the help of independent variable. Since, we will try to fit a straight line to know the relationship (that's why this is called linear regression), we should first see whether a linear relationship exists before trying to find a linear relationship. The best way to SEE this is to plot a scatterplot. If the dots seem to represent a line, then only regression should be done. However, if they seem to make a curve,

exponential etc. we shouldn't fit the linear regression and try to fit the curve equation depending on what we think best represents the relationship.

The linear regression equation can be written as:

$Y = a + bX + e$

Y = dependent variable

X = independent variable

a = constant

e = Error term

When we are predicting the value of dependent variable Y after knowing the values of X, we actually don't expect that we will be able to know the dependent variable fully if we know the independent. The error term signifies the amount of variation not explained by independent variable. In bivariate regression, the coefficient of X is nothing but the correlation coefficient only. Let's take an example. Suppose we want to predict sales after increasing advertising spending so the relationship is

Sales (in Rs.)= constant + b (Advertising spending in Rs.) + Error term

Our hypothesis here is that we cannot predict sales by advertising, which means, there is no relationship between the two, which means whether we increase or decrease Advertising spending Sales will not change which essentially means b=0.

If b=0, then any value of Advertising spending will be multiplied by 0 and will not affect sales. So

H0 : b=0 and

Ha: $b \neq 0$

Recall that in bivariate regression, b is nothing but correlation coefficient and in doing correlation also we tested this hypothesis. Therefore, it clarifies that until and unless, correlation is significant, there is no point in doing regression.

The data will be entered like this

Observation	Sales in Rs.	Advertising spending in Rs.
1	32	45
2	25	29
3	86	56
4	45	76

To do regression,

Go to SPSS/Analyse/Regression/Linear

Put the dependent variable in the dependent box and the independent in the independent and click OK. To interpret the output, first see the ANOVA table. If the P value in ANOVA is less than .05, then the hypothesis that b=0, is rejected and we conclude that b≠0 and hence the independent effects the dependent variable.

To know how much variation in Y can be explained/accounted for by the independent X, look at the R^2 value in the model summary.

To find the equation, look at the unstandardized coefficients under heading B.

The equation is Y (Dependent) = the first entry below B + the second entry in the B column * independent

Recall that we earlier said that the objective of regression is to know the value of Y if we know X. So we can put any value of X in the above equation and calculate the corresponding value of Y.

However, to predict Y when we know X, is not the only objective in regression especially when we have multiple predictors (independents) of one dependent. For example, the sales of a store may be dependent on the amount of advertising, number of service lines, total number of assortments, total time the store is open etc. Observe that I have taken all variables as metric because we have already defined that in correlation and regression we consider the variables as metric. For simplicity, I will take only two independents as the method is exactly the same whether there are two or more than two independents. So

Y (Sales) = a (constant) + b (advertising) + c (total time the store is open) + Error

The null hypothesis again is that sales cannot be predicted by advertising and service lines. Again the hypothesis is

H0: b=c=0

Ha: At least one of the b,c is non-zero. Now to answer what's the value of Sales, when we know advertising and total time the store is open, we will go to

SPSS/Analyse/Regression/Linear and this time put both independents in the independent box and press OK.

The output has to be exactly read as earlier. First we see the ANOVA. If anova is not significant, i.e., p value is greater than .05, the null hypothesis has to be accepted, and no further conclusion can be made except that the two independents don't affect sales.

If the P value is less than .05, at least one of them effects so three cases are possible. First, advertising effects but not total time, second, total time effects but not advertising, third both effect.

The R^2 value has to be interpreted as before only thing is that it is the combined explanation of sales by both variables taken together.

The equation will have to be formed again in the similar manner. To know which of the two variables effect, we can look at the t-statistic and the p value associated with the t-statistic. If the p value is less than .05, that independent affects the dependent variable otherwise not.

Now, let's ask another managerial question. The manager asks, I can change only one thing, either advertising budget or number of hours the store is open, which one I should change. Essentially, he is asking us to compare the two independents and say which one affects more. We are assuming both effect sales which were found using t statistic. Let's take a hypothetical value

Sales = 23 + 8 * advertising budget + 5 * total time

On simple intuition, we can say that the coefficient of advertising is 8 so one unit change in advertising will bring 8 units of change in sales whereas one unit change in total time will bring 5 units of change in sales, so advertising effects more.

However, there is flaw in this.

Advertising budget is measured in Rs. and total time is measured in say hours, we cannot actually compare Rupees with say hours as the two are measured in different units. Recall, 1kg=1000gms which means if you decrease the unit of measurement from kg to gm, the coefficient increases to make the two sides equal. Similarly, if the advertising budget is measured in dollars instead of Rupees, the coefficient 8 would become one 60^{th} time of 8 to keep the two expressions equal (assumption 1dollar=60Rs) which essentially means total time (by previous logic will have a higher coefficient) and the conclusion will change.

Simply put, we can't compare two things until and unless they are in the same unit or else UNITLESS. Recall, that you learned something called standardization, by which you used to convert any normal distribution to corresponding standard normal distribution. Yes, the process is simple. You subtract mean from each value so that the new mean becomes 0. You divide each value by standard deviation so that the new standard deviation becomes 1. Wow, you have got standard normal distribution with mean 0 and standard deviation as 1.

Just for simplicity, we will subtract each mean of advertising budget from each advertising budget value and divide it with standard deviation of the advertising budget. The new variable will be standardized.

Let's see the effect of standardization:

Z =(x – mu)Rupees/standard deviation in Rupees (Rupees Rupees cancel out in numerator and denominator)

The numerator is in Rs as both x and mu are in Rs. The denominator is also in Rs.as it is the standard deviation. So the resultant standardized variable is free from any measurement unit.

Simply put to compare the effect of two or more dependents we have to see the standardized coefficients given below Beta to answer which of the two variables affect sales more. Notice, there is no constant in standardized equation, why?