

## 23. Multiple Regression

In many real-world situations researchers will have several independent variable. Spreadsheets for examples are here.

### 23.1. Example.

The human resources director for a chain of car dealers is interested in the attributes that influence sales. She randomly selects twenty sales people employed by the dealership and records their sales for the month of April, their scores on a standardized IQ test, and their scores on a standardized test for extroversion. She obtains the following results:

Sales	IQ	Extroversion Scale
\$2,625	89	21
\$2,700	93	24
\$3,100	91	21
\$3,150	122	23
\$3,175	115	27
\$3,100	100	18
\$2,700	98	19
\$2,475	105	16
\$3,625	112	23
\$3,525	109	28
\$3,225	130	20
\$3,450	104	25
\$2,425	104	20
\$3,025	111	26
\$3,625	97	28
\$2,750	115	29
\$3,150	113	25
\$2,600	88	23
\$2,525	108	19
\$2,650	101	16

The researcher plans to use this information to rate applicants for sales jobs. If she has an applicant with an IQ of 110 and a score of 23 on the extrovert scale, what sales would she predict based on this data?

**Solution.** The MultipleRegression tab of the AnalyzeThis spreadsheet answers these questions and more. First we need to identify which variables are independent and which is dependent. In this example, the HR director is interested in what influences car sales, so this must be the dependent variable. The independent variables are then IQ and extroversion.

The model that the HR director proposes is that there is a linear relationship between sales, IQ, and extroversion:

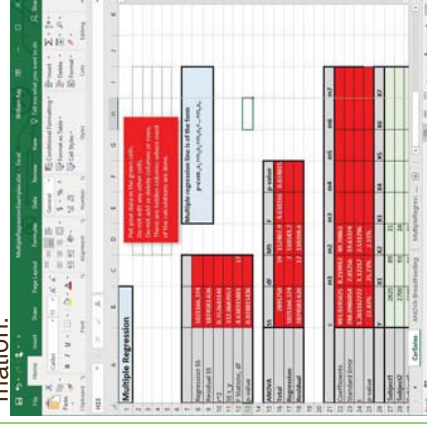
$$y = c + m_1x_1 + m_2x_2$$

where

$y =$	car sales
$x_1 =$	IQ score
$x_2 =$	extroversion score

You can then enter the data into the cells B27:D46 of the spreadsheet. Note that  $y$  goes into the first column and the other two scores into the next two columns.

The analysis position of the spreadsheet gives you quite a bit of information.



The first section gives the regression statistics for testing

$$H_0 : r^2 = 0 \quad \text{against}$$

$$H_A : r^2 > 0$$

The key value is the  $p$ -value of 0.0248 or 2.48%, which means that we have significant but not highly significant evidence in support of  $H_A$ . From this we believe that the observed value of  $r^2$  cannot be attributed to chance.

The second segment, labeled **ANOVA** we can skip for now.

The final section gives you the values of  $c$ ,  $m_1$  and  $m_2$  in the above model, and so

$$y = 993.92 + 9.219x_1 + 49.70x_2.$$

From this, it's easy to substitute in  $x_1 = 110$  and  $x_2 = 23$  to predict sales of \$3,341 for the applicant with an IQ score of 110 and an extroversion score of 23.

But there is some additional information. For example, there are  $p$ -values for  $m_1$  and  $m_2$ . These relate to the hypotheses

$$H_1 : m_1 \neq 0 \quad \text{and} \quad H_2 : m_2 \neq 0.$$

Thus, if we believe  $m_2$  is not zero, the chance we are wrong is 2.15%. On the other hand, if we believe  $m_1$  is not zero, the chance we are wrong is 25%.

What does this say about using the model to predict sales?

In particular, this means we should not use  $m_1$  to predict sales, since we cannot assume its value is nonzero. Since we can't use  $m_1$ , that means that the above prediction of \$3,341 is also not reliable, since it used  $m_1$ . This suggests running another ANOVA using just extroversion and sales and omitting the variable IQ.

On the hand, the earlier  $p$ -value for  $r^2$  lets us conclude that there is a connection between the variables. What the information on the coefficients means is that the connection, while real, is not strong enough to use for prediction.

The `MultipleRegression` tab of `AnalyzeThis` has many powerful features built into it, and tests more than one hypothesis. Multiple regression can even provide an alternative way of thinking about ANOVA.

### 23.2. Example.

There is a folk legend that if a mother drinks a beer prior to nursing her infant, the child will take in more breast milk. To test this, a nurse working in the maternity ward of a hospital randomly selected 40 nursing mothers and randomly divided them into four groups as follows:

- Group I received instruction on breast feeding and ingested 10 oz of beer prior to nursing;
- Group II received the instruction and ingested 10 oz of a non-alcoholic beverage prior to nursing;
- Group III received instruction but was offered no beverage prior to nursing;
- group IV received neither instruction nor beverage prior to nursing.

The researcher then weighed the infants before and after nursing and recorded the difference in weight, those differences being the amount ingested.

**Solution.** Using the ANOVA tab of `AnalyzeThis`...

Regression ANOVA - Excel

You may have unbalanced sample sizes, i.e., different numbers in each treatment group. This spreadsheet uses the method of least squares, which is appropriate for continuous and generally continuous as opposed to discrete.

For completeness, this spreadsheet includes both the analysis in Table I and in Table II. Table II contains information on "sums of squares" and "mean square" statistics, but gives the same conclusions as Table I.

Tests the null hypothesis that all the column means are the same against the alternative that they are not.

Treatment Groups	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8
Mean	4.62	3.59	3.69	3.89				
Sample Size	10	10	10	10				

Value	Degrees of Freedom
0.749144915	19
0.102648915	1
0.939995	36
1.927925171	
14.25%	

SS	df	MSE	F	p	Critical Value
29.66975	39	0.760763	3.07975	14.25%	2.8663
0.0665	36	0.001847			
0.00975	1	0.00975			

ANOVA	df	MSE	F	p-value
Total	29.66975	39	3.07975	0.0005
Regression	4.19025	3	3.369917	0.0305
Residual	25.4795	36	0.710491	

Coefficients	beta_0	beta_1	beta_2	beta_3	beta_4	beta_5	beta_6	beta_7
Standard Error	0.26677489	0.176951	0.176951	0.176951	0.176951	0.176951	0.176951	0.176951
t Stat	14.98803119	1.981901	-0.79897	-0.31077	0.9695	-0.1276	-0.1276	-0.1276
P-value	0.00000000	0.04876	0.42876	0.75424	0.33424	0.33424	0.33424	0.33424

We can also analyze the data using multiple regression by way of indicator variables:

$$m_1 = \begin{cases} 1 & \text{if the observation is in Group I} \\ 0 & \text{otherwise} \end{cases}$$

$$m_2 = \begin{cases} 1 & \text{if the observation is in Group II} \\ 0 & \text{otherwise} \end{cases}$$

$$m_3 = \begin{cases} 1 & \text{if the observation is in Group III} \\ 0 & \text{otherwise} \end{cases}$$

An observation is in Group IV exactly when

$$m_1 = m_2 = m_3 = 0$$

so we don't need an indicator variable for this group. Using MultipleRegression gives identical results to ANOVA.

Multiple Regression

Multiple regression line is of the form  $y = \beta_0 + \beta_1 m_1 + \beta_2 m_2 + \beta_3 m_3 + \beta_4 m_4$ .

Multiple regression line is of the form  $y = \beta_0 + \beta_1 m_1 + \beta_2 m_2 + \beta_3 m_3 + \beta_4 m_4$ .

Regression SS	df	MSE	F	p-value
4.19025	3	1.39675	3.369917	0.0305
25.4795	36	0.710491		
Total	29.66975	39	3.07975	0.0005

Coefficients	beta_0	beta_1	beta_2	beta_3	beta_4	beta_5	beta_6	beta_7
Standard Error	0.26677489	0.176951	0.176951	0.176951	0.176951	0.176951	0.176951	0.176951
t Stat	14.98803119	1.981901	-0.79897	-0.31077	0.9695	-0.1276	-0.1276	-0.1276
P-value	0.00000000	0.04876	0.42876	0.75424	0.33424	0.33424	0.33424	0.33424

Notice that the ANOVA section in the spreadsheet replicates exactly the table from the ANOVA tab that we did earlier. In addition, the  $F$ -statistic is has the same value as the one that tests

$$H_0 : r^2 = 0; \text{ against}$$

$$H_A : r^2 > 0$$

Since the  $p$ -value is 0.1423 or 14.25%, we reject  $H_0$  and believe that the value of  $r^2$  cannot be attributed to chance. Similarly, the ANOVA statistic is telling us that the differences in the means cannot be attributed to chance. The approach using linear regression with indicator variables is thus seen to be statistically equivalent to the test comparing means.