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:

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?

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

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:

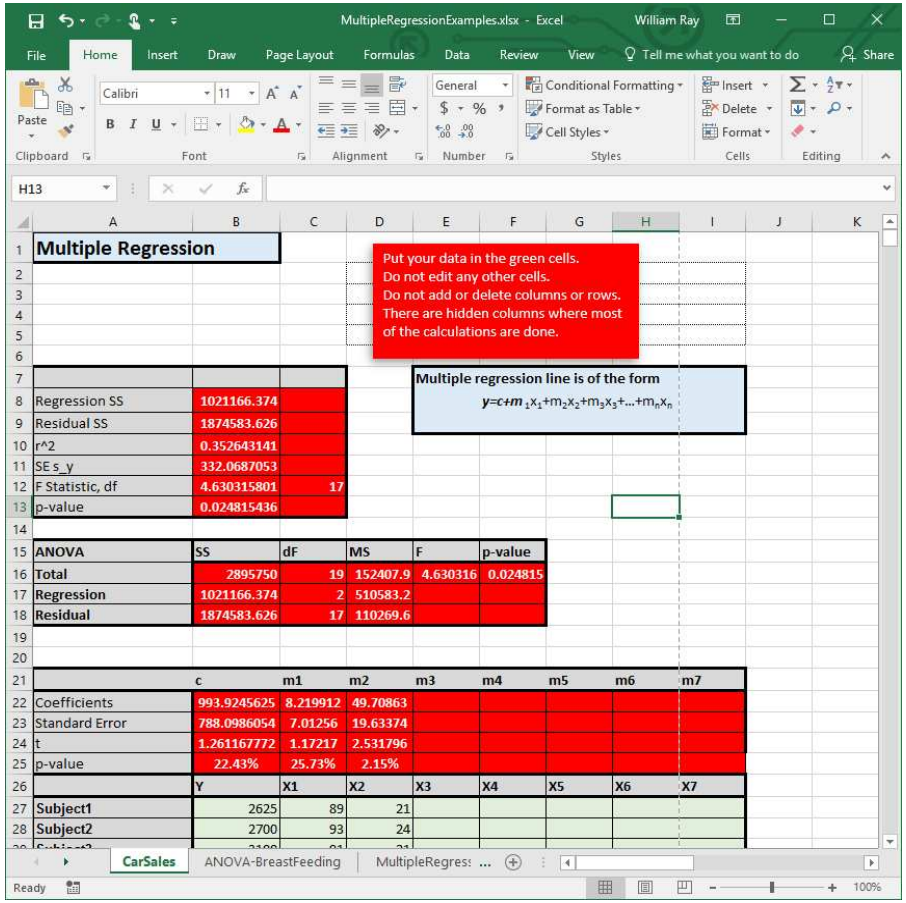
$$y = c + m_1 x_1 + m_2 x_2$$

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.



MultipleRegressionExamples.xlsx - Excel — William Ray

	A	B	C	D	E	F	G	H	I	J	K
1	**Multiple Regression**										
2											
3											
4											
5											
6											
7											
8	Regression SS	1021166.374									
9	Residual SS	1874583.626									
10	r^2	0.352643141									
11	SE s_y	332.0687053									
12	F Statistic, df	4.630315801	17								
13	p-value	0.024815436									
14											
15	**ANOVA**	**SS**	**dF**	**MS**	**F**	**p-value**					
16	**Total**	2895750	19	152407.9	4.630316	0.024815					
17	**Regression**	1021166.374	2	510583.2							
18	**Residual**	1874583.626	17	110269.6							
19											
20											
21		**c**	**m1**	**m2**	**m3**	**m4**	**m5**	**m6**	**m7**		
22	Coefficients	993.9245625	8.219912	49.70863							
23	Standard Error	788.0986054	7.01256	19.63374							
24	t	1.261167772	1.17217	2.531796							
25	p-value	22.43%	25.73%	2.15%							
26		**Y**	**X1**	**X2**	**X3**	**X4**	**X5**	**X6**	**X7**		
27	**Subject1**	2625	89	21							
28	**Subject2**	2700	93	24							

Put your data in the green cells.
Do not edit any other cells.
Do not add or delete columns or rows.
There are hidden columns where most of the calculations are done.

Multiple regression line is of the form

$y=c+m_1x_1+m_2x_2+m_3x_3+\ldots+m_nx_n$

The first section gives the regression statistics for testing

$$\begin{aligned} \boldsymbol{H_0}: &\quad \boldsymbol{r^2} = 0 \quad \text{against} \\ \boldsymbol{H_A}: &\quad \boldsymbol{r^2} > 0 \end{aligned}$$

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 $\boldsymbol{H_A}$. From this we believe that the observed value of $\boldsymbol{r^2}$ cannot be attributed to chance.
The second segment, labeled ANOVA we can skip for now.
The final section gives you the values of $\boldsymbol{c}$, $\boldsymbol{m_1}$ and $\boldsymbol{m_2}$ in the above model, and so

$$\boldsymbol{y} = 993.92 + 9.219\boldsymbol{x_1} + 49.70\boldsymbol{x_2}.$$

From this, it's easy to substitute in $\boldsymbol{x_1} = 110$ and $\boldsymbol{x_2} = 23$ 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.

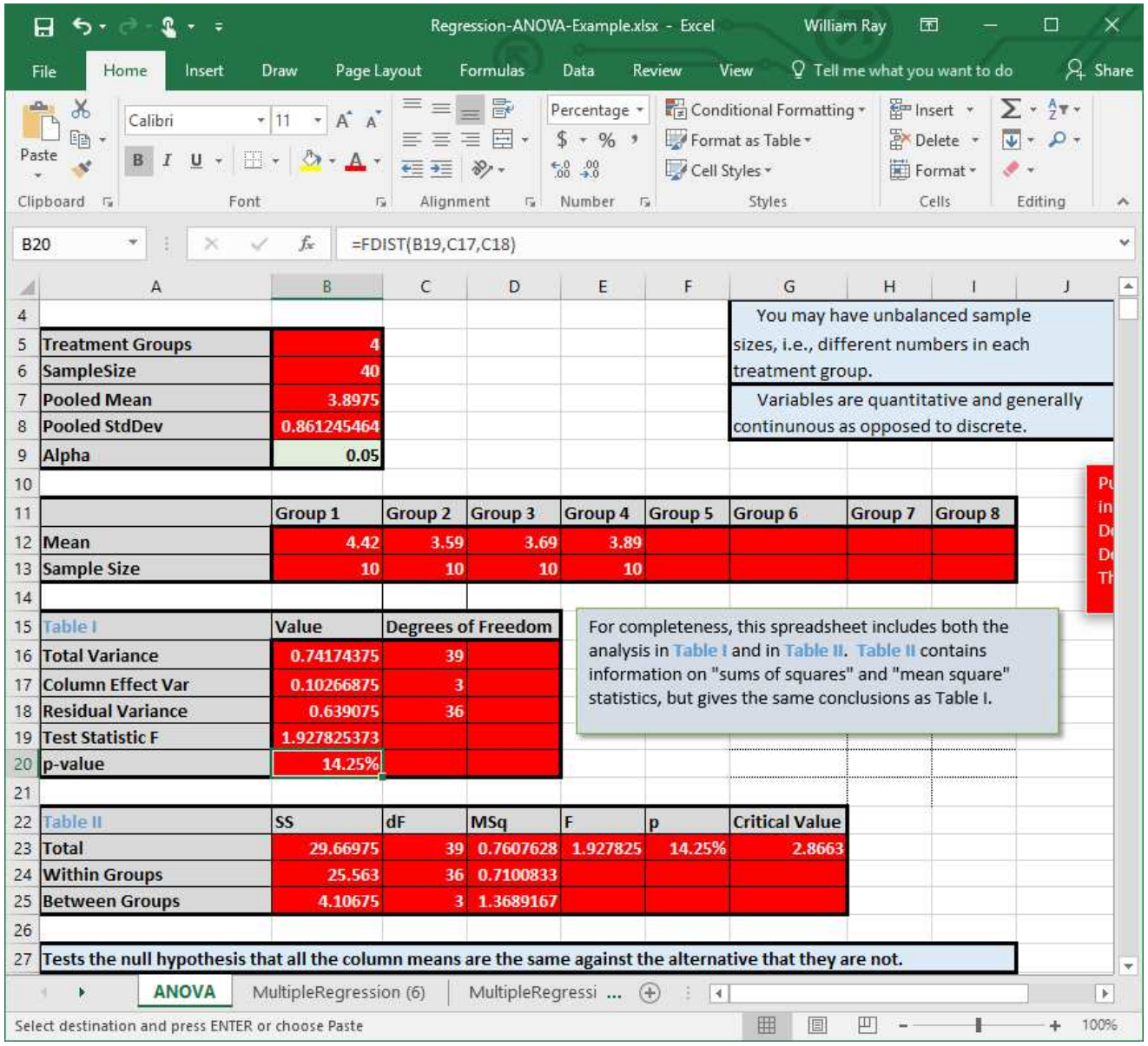
23.2. Example.

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`...



Treatment Groups	4	
SampleSize	40	
Pooled Mean	3.8975	
Pooled StdDev	0.861245464	
Alpha	0.05	

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8
Mean	4.42	3.59	3.69	3.89				
Sample Size	10	10	10	10				

Table I	Value	Degrees of Freedom
Total Variance	0.74174375	39
Column Effect Var	0.10266875	3
Residual Variance	0.639075	36
Test Statistic F	1.927825373	
p-value	14.25%	

Table II	SS	dF	MSq	F	p	Critical Value
Total	29.66975	39	0.7607628	1.927825	14.25%	2.8663
Within Groups	25.563	36	0.7100833			
Between Groups	4.10675	3	1.3689167			

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

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

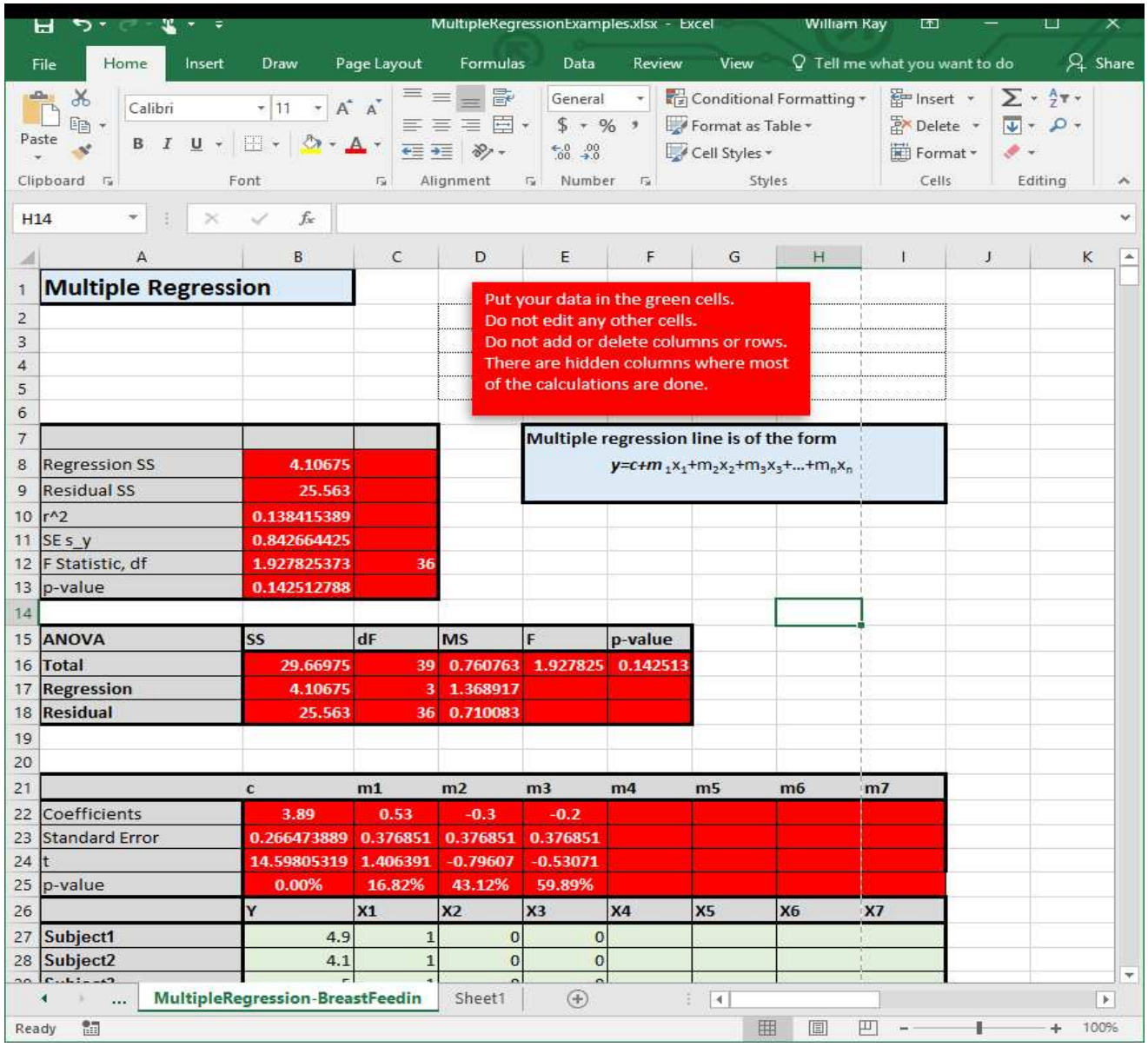
$$m_2 = \begin{cases} 1 & \text{if the observation is in Group II} \\ 0 & \text{otherwise} \end{cases}$$

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

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`.


MultipleRegressionExamples.xlsx - Excel
William Ray
File
Home
Insert
Draw
Page Layout
Formulas
Data
Review
View
Tell me what you want to do
Share
Paste
Calibri
11
General
Conditional Formatting
Format as Table
Cell Styles
Insert
Delete
Format
Clipboard
Font
Alignment
Number
Styles
Cells
Editing
H14
Multiple Regression
Put your data in the green cells.
Do not edit any other cells.
Do not add or delete columns or rows.
There are hidden columns where most of the calculations are done.
Multiple regression line is of the form
y=c+m1x1+m2x2+m3x3+...+mnxn
Regression SS 4.10675
Residual SS 25.563
r^2 0.138415389
SE s_y 0.842664425
F Statistic, df 1.927825373 36
p-value 0.142512788
ANOVA SS dF MS F p-value
Total 29.66975 39 0.760763 1.927825 0.142513
Regression 4.10675 3 1.368917
Residual 25.563 36 0.710083
c m1 m2 m3 m4 m5 m6 m7
Coefficients 3.89 0.53 -0.3 -0.2
Standard Error 0.266473889 0.376851 0.376851 0.376851
t 14.59805319 1.406391 -0.79607 -0.53071
p-value 0.00% 16.82% 43.12% 59.89%
Y X1 X2 X3 X4 X5 X6 X7
Subject1 4.9 1 0 0
Subject2 4.1 1 0 0
MultipleRegression-BreastFeedin
Sheet1
Ready
100%

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

$$\begin{aligned} H_0: &\quad r^2 = 0; \quad \text{against} \\ H_A: &\quad r^2 > 0 \end{aligned}$$

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.