APPENDIX 16.2: MULTIPLE REGRESSION IN EXCEL

SPSS is a powerful tool for doing statistical analyses, but Excel has functions that do many of the same things. The **LINEST** function in Excel allows us to perform multiple regression rather easily. The syntax for the **LINEST** function is

LINEST(y,x_1 . . . x_k,[const],[stats]).

The outcome variable *y* is a column of numbers and $x_1 \ldots x_k$ is one or more columns of predictor variables. The optional argument *const* takes values of 0 and 1 (or **FALSE** and **TRUE**). (Optional arguments are always enclosed in square parentheses in Excel.) When *const* is set to 0, the regression equation is computed assuming that the intercept is 0; this makes little sense for most of our applications. When *const* is set to 1, the regression equation is computed in the usual way. The final argument, *stats*, also takes values of 0 and 1 (or **FALSE** and **TRUE**). When *stats* = 0, then only the coefficients of the regression equation are returned. When *stats* = 1, then the coefficients of the regression equation are returned along with a number of other statistics that we'll discuss below.

LINEST is different from all other functions we've discussed before because it returns more than a single value. In the language of Excel, **LINEST** is called an *array function*, and extra steps must be taken to have multiple values returned. Figure 16.A2.1 shows how we enter arguments into the **LINEST** function (see cell **E8**). The *y* values are in cells **A2:A21**, and the three predictors $x_1 \ldots x_k$ are in cells **B2:D21**. *const* is set to 1 and *stats* is set to 1. When we press return, a single number appears in cell **E8**, as shown in Figure 16.A2.2.

The number in cell **E8** is the regression coefficient for the third predictor; i.e., b_3. Before we look at *how* to obtain the remaining coefficients and statistics, we will look at how these will be arranged. The cells in the array **E2** to **H6** have text labels that correspond to the way that Excel returns the coefficients and statistics of the regression analysis. The top-left element will be the regression coefficient for the last, or *k*th, predictor; i.e., b_k. In our example, there are three predictors, so $k = 3$. As we move from left to right in this row, we see the remaining regression coefficients (b_2 and b_1 in this case) and the intercept (*a*). Below each of these coefficients will be their standard errors (s_{b_k}, $s_{b_{k-1}}, \ldots, s_{b_1}$ and s_a). The text in the remaining cells (**E4:F6**) should be self-explanatory.

FIGURE 16.A2.1 ■ Entering the LINEST Function

LINEST =LINEST(A2:A21,B2:D21,1,1)

	A	B	C	D	E	F	G	H
1	y	x_1	x_2	x_3				
2	5.18	4.40	4.80	4.26	b_3	b_2	b_1	a
3	4.85	3.45	3.01	4.43	s_{b3}	s_{b2}	s_{b1}	s_a
4	5.92	5.24	4.96	5.58	R^2	s_{est}		
5	5.90	5.82	6.88	6.14	F	df_{Regr}		
6	6.08	4.85	5.90	5.70	ss_{Regr}	ss_{Error}		
7	5.07	2.97	4.13	3.67				
8	4.22	5.68	4.58	5.58	=LINEST(A2:A21,B2:D21,1,1)			
9	5.81	5.30	5.45	5.75	LINEST(known_y's, [known_x's], [const], [stats])			
10	4.14	4.09	4.76	4.26				
11	5.41	4.40	4.48	3.88				
12	6.79	6.24	5.35	5.58				
13	3.42	3.58	2.57	4.08				
14	4.54	3.19	3.57	2.66				
15	4.72	4.35	5.52	5.41				
16	4.91	6.49	4.64	6.34				
17	4.26	3.21	3.12	3.04				
18	4.73	4.25	6.02	4.57				
19	1.83	3.04	3.91	3.32				
20	4.24	3.82	4.80	5.09				
21	4.59	4.52	2.81	4.11				
22								

FIGURE 16.A2.2 ■ Initial LINEST Output

E8 =LINEST(A2:A21,B2:D21,1,1)

	A	B	C	D	E	F	G	H
1	y	x_1	x_2	x_3				
2	5.18	4.40	4.80	4.26	b_3	b_2	b_1	a
3	4.85	3.45	3.01	4.43	s_{b3}	s_{b2}	s_{b1}	s_a
4	5.92	5.24	4.96	5.58	R^2	s_{est}		
5	5.90	5.82	6.88	6.14	F	df_{Regr}		
6	6.08	4.85	5.90	5.70	ss_{Regr}	ss_{Error}		
7	5.07	2.97	4.13	3.67				
8	4.22	5.68	4.58	5.58	-0.130			
9	5.81	5.30	5.45	5.75				
10	4.14	4.09	4.76	4.26				
11	5.41	4.40	4.48	3.88				
12	6.79	6.24	5.35	5.58				

To obtain the remaining coefficients and statistics, we highlight an array of cells having five rows and $k + 1$ columns, with the top-left cell of this array containing the result returned by **LINEST**. This is shown as the shaded rectangle in Figure 16.A2.3. With this region highlighted, we enter a sequence of keystrokes that will differ depending on whether you are using a Macintosh or a PC. On a Macintosh, you first enter <control>u. That is, you press the control key and the u key simultaneously. This will highlight the *y* and *x*

FIGURE 16.A2.3 ■ Specifying the Output Cells

E8 =LINEST(A2:A21,B2:D21,1,1)

	A	B	C	D	E	F	G	H
1	y	x_1	x_2	x_3				
2	5.18	4.40	4.80	4.26	b_3	b_2	b_1	a
3	4.85	3.45	3.01	4.43	s_{b3}	s_{b2}	s_{b1}	s_a
4	5.92	5.24	4.96	5.58	R^2	s_{est}		
5	5.90	5.82	6.88	6.14	F	df_{Regr}		
6	6.08	4.85	5.90	5.70	ss_{Regr}	ss_{Error}		
7	5.07	2.97	4.13	3.67				
8	4.22	5.68	4.58	5.58	-0.130			
9	5.81	5.30	5.45	5.75				
10	4.14	4.09	4.76	4.26				
11	5.41	4.40	4.48	3.88				
12	6.79	6.24	5.35	5.58				

To obtain the remaining coefficients and statistics from the regression analysis, we highlight an array of five rows and $k + 1$ columns, with the result returned by **LINEST** in the top left. This region is shown as the shaded rectangle.

FIGURE 16.A2.4 ■ Preparing the Full Analysis

F.INV.RT =LINEST(A2:A21,B2:D21,1,1)

	A	B	C	D	E	F	G	H
1	y	x_1	x_2	x_3				
2	5.18	4.40	4.80	4.26	b_3	b_2	b_1	a
3	4.85	3.45	3.01	4.43	s_{b3}	s_{b2}	s_{b1}	s_a
4	5.92	5.24	4.96	5.58	R^2	s_{est}		
5	5.90	5.82	6.88	6.14	F	df_{Regr}		
6	6.08	4.85	5.90	5.70	ss_{Regr}	ss_{Error}		
7	5.07	2.97	4.13	3.67				
8	4.22	5.68	4.58	5.58	=LINEST(A2:A21,B2:D21,1,1)			
9	5.81	5.30	5.45	5.75				
10	4.14	4.09	4.76	4.26				
11	5.41	4.40	4.48	3.88				
12	6.79	6.24	5.35	5.58				

When the cursor is placed in the formula bar, we say that the **LINEST** cell has been selected. (This can also be accomplished on a Macintosh by pressing <control>u.) Once this is done, press ⌘⏎ on a Macintosh, or press <control><shift>enter on a PC.

FIGURE 16.A2.5 ■ LINEST Regression Output

L28

	A	B	C	D	E	F	G	H
1	y	x_1	x_2	x_3				
2	5.18	4.40	4.80	4.26	b_3	b_2	b_1	a
3	4.85	3.45	3.01	4.43	s_{b3}	s_{b2}	s_{b1}	s_a
4	5.92	5.24	4.96	5.58	R^2	s_{est}		
5	5.90	5.82	6.88	6.14	F	df_{Regr}		
6	6.08	4.85	5.90	5.70	ss_{Regr}	ss_{Error}		
7	5.07	2.97	4.13	3.67				
8	4.22	5.68	4.58	5.58	-0.130	0.298	0.536	1.698
9	5.81	5.30	5.45	5.75	0.417	0.240	0.374	0.965
10	4.14	4.09	4.76	4.26	0.428	0.885	#N/A	#N/A
11	5.41	4.40	4.48	3.88	3.983	16.000	#N/A	#N/A
12	6.79	6.24	5.35	5.58	9.360	12.534	#N/A	#N/A

Cells **E2** to **H6** denote the quantities in cells **E8** to **H12**.

columns of data, as shown in Figure 16.A2.4. The same thing can be accomplished by clicking in the formula bar at the top of the sheet. If you do this, there is no need to enter <control>u. The next step is to press the command key (⌘) and then return (⏎). When this is done, the regression coefficients and statistics will fill in the highlighted region, as shown in Figure 16.A2.5. On a PC, you first place the cursor in the formula bar at the top and then press <control><shift>enter. The result will be the same as in Figure 16.A2.5.

The text in cells **E2** to **H6** denotes the quantities in cells **E8** to **H12**. The first row (**E8** to **H8**) shows the regression coefficients, and the second row (**E9** to **H12**) shows the standard errors of the regression coefficients. Cells **E10** and **F10** contain R^2 and s_{est}, respectively. Cells **E11** and **F11** contain F_{obs} and $df_{regression}$, respectively. Cells **E12** and **F12** contain $ss_{regression}$ and ss_{error}, respectively. The **#N/A** symbol appears in cells for which **LINEST** doesn't return a value.

Figure 16.A2.6 shows the regression analysis performed by SPSS on the data shown in Figure 16.A2.1. A quick comparison between this and Figure 16.A2.5 shows that exactly the same regression coefficients, standard errors, R^2, s_{est}, F_{obs}, $ss_{regression}$, and ss_{error} are obtained in both cases. Excel does not compute the *t*-statistics and *p*-values associated with the regression coefficients, but this can be done easily in Excel by dividing each b_i by its estimated standard error, s_{b_i}. The *p*-value can be obtained using the **T.DIST** function discussed in earlier appendices. There are $n - k - 1$ degrees of freedom associated with each of these tests.

We did not discuss confidence intervals around the regression coefficients in Chapter 16, but these can be easily computed as

$$b_i \pm t_{\alpha/2}(s_{b_i}),$$

where $t_{\alpha/2}$ is based on $n - k - 1$ degrees of freedom. Therefore, the 95% confidence interval around b_1 would be

$$\text{CI} = b_i \pm t_{\alpha/2}(s_{b_i})$$
$$= 0.536 \pm 2.12(0.374) = [-0.26,\ 1.33].$$

We will discuss the regression coefficients in more detail in Chapter 17.

To compute the *p*-value associated with F_{obs}, we use the **F.DIST** function in Excel, as shown in Figure 16.A2.7. The **F.DIST** function has the following syntax:

F.DIST(F_{obs},$df_{regression}$,df_{error},cumulative).

FIGURE 16.A2.6 ■ SPSS Regression Output

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	x3, x2, x1[b]	.	Enter

a. Dependent Variable: y

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.654[a]	.428	.320	.88508

a. Predictors: (Constant), x3, x2, x1

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9.360	3	3.120	3.983	.027[b]
	Residual	12.534	16	.783		
	Total	21.894	19			

a. Dependent Variable: y

b. Predictors: (Constant), x3, x2, x1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.698	.965		1.760	.098
	x1	.536	.374	.536	1.434	.171
	x2	.298	.240	.319	1.243	.232
	x3	−.130	.417	−.129	−.313	.759

a. Dependent Variable: y

FIGURE 16.A2.7 ■ The F.DIST Function in Excel

b_3	b_2	b_1	a
s_{b3}	s_{b2}	s_{b1}	s_a
R^2	s_{est}		
F	df_{Regr}		
SS_{Regr}	SS_{Error}		
-0.13028	0.298012	0.535831	1.69792
0.416701	0.239674	0.373603	0.964898
0.427522	0.885077	#N/A	#N/A
3.982889	16	#N/A	#N/A
9.360119	12.53378	#N/A	#N/A
n	20	=COUNT(A2:A21)	
df_{total}	19	=F14-1	
$df_{regr.}$	3		
df_{error}	16	=F15-F16	
P(F)	0.973	=F.DIST(E11,F16,F17,1)	
p	0.027	=1-F18	

When *cumulative* is set to 1 (as in the second-to-last row in Figure 16.A2.7), **F.DIST** returns the proportion of the *F*-distribution with $df_{regression}$ and df_{error} falling below F_{obs}. In this example, the proportion of an *F*-distribution with 3 and 16 falling below F_{obs} = 3.983 is .973. The proportion above F_{obs} = 3.983 is 1 − .973 = .027, as shown in the last row of Figure 16.A2.7. Therefore, the three-predictor model explains a statistically significant proportion of the variability in *y*.

Note that the **F.DIST.RT** function

$$\textbf{F.DIST.RT}(F_{obs},df_{regression},df_{error})$$

accomplishes the same thing by returning the proportion of the *F*-distribution above F_{obs}; i.e., the proportion of the *F*-distribution in the right tail (RT).

APPENDIX 16.3: POWER FOR R^2 AND ΔR^2

The small data set discussed in Chapter 16 was chosen to illustrate the fundamentals of standard multiple regression analysis. Missing from our discussion of the data set was the notion of power. We've seen before that a prospective power analysis should be conducted before data collection. Of course, we could not have discussed power at the outset because we have to know something about multiple regression before we can discuss power. Now that we are in a position to discuss power, we will work through the kind of analyses that should be conducted before data collection.

As we've seen, researchers almost universally use significance testing to assess the fit of a regression model to data (the omnibus *F* for R^2) or the change in explained variance (ΔR^2) resulting from the addition of one or more predictors to a model. We can assess power for both levels of analysis using G*Power.

The measure of effect size used for the omnibus test is Cohen's f^2, which is defined as

$$f^2 = \frac{P^2}{1 - P^2}. \qquad (16.A3.1)$$

where P^2 is the population version of R^2 that you wish to detect in a significance test. Equation 16.A3.1 is essentially the first term of F_{obs} from equation 16.8. To conduct a power analysis, we simply decide what value of P^2 would be meaningful, compute f^2, and then enter this in G*Power as shown in Figure 16.A3.1.

If we were to assume that $P^2 = .53$ is the explained variance that would be interesting in a given research setting, then equation 16.A3.1 shows that this corresponds to $f^2 = P^2/(1 - P^2) = .53/.47 = 1.127$. In G*Power, we choose F-tests from the Test family drop-down list and Linear multiple regression: Fixed model, R2 deviation from zero from the Statistical test drop-down list. From the Type of power analysis drop-down list, we choose A priori: Compute required sample size - given α, power, and effect size. $f^2 = 1.127$ is entered in the Effect size box, and α, power, and number of predictors are entered in the boxes below this. When the analysis has been described, clicking Calculate returns the result of the analysis in the Output parameters area of the dialog.

Figure 16.A3.1 shows that we require 13 subjects to detect $P^2 = .53$ with power = .8, when there are two predictors and α = .05. (In fact, we would actually achieve power = .84.) This result is probably not surprising; $P^2 = .53$ is quite a large proportion of explained variance and should therefore be easy to detect.

The more interesting case is determining the sample size required to detect a specific change in explained variance. The measure of effect size used for power analyses related to ΔR^2 is a second version of Cohen's f^2, defined as

$$f^2 = \frac{P^2_{larger} - P^2_{smaller}}{1 - P^2_{larger}} = \frac{\Delta P^2}{1 - P^2_{larger}}. \qquad (16.A3.2)$$

Equation 16.A3.2 is the first term of F_{change} from equation 16.10. However, we are considering parameters and not statistics, so we replace R with P.

To conduct this analysis, one must have some sense of how much variance will be explained by our initial or reduced set of predictors, which we described earlier as the smaller model. So, we must have an idea of what $P^2_{smaller}$ would be. Such estimates could be drawn from previous research. In addition, we would have to know what increase in explained variance (ΔP^2) would be meaningful in the context of our research

FIGURE 16.A3.1 ■ A Power Analysis for P^2

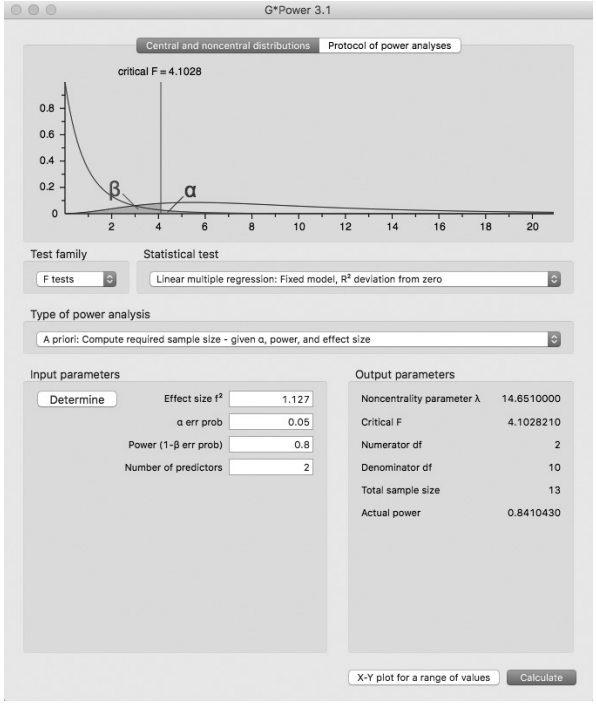
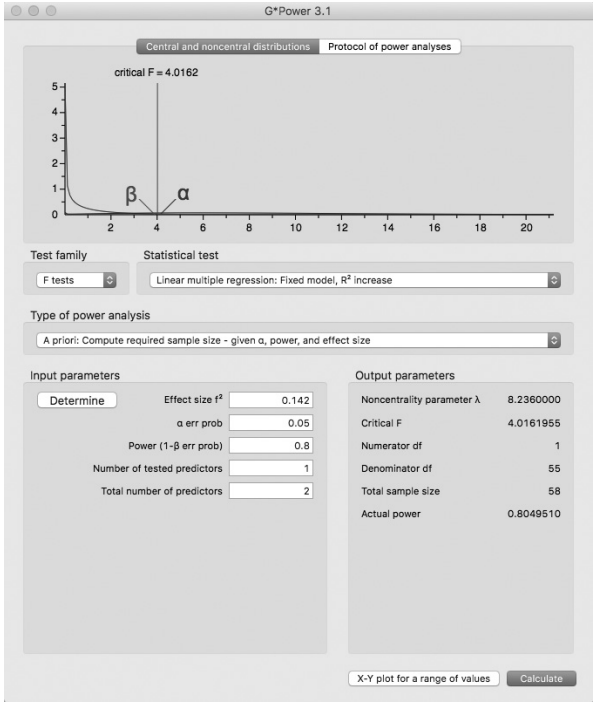


FIGURE 16.A3.2 ■ A Power Analysis for ΔP^2



question. With these questions answered, we can say that $P^2_{larger} = P^2_{smaller} + \Delta P^2$ and then determine f^2 from equation 16.A3.2.

If we were to expect $P^2_{smaller}$ to be .47 and judge that $\Delta P^2 = .066$ would be a meaningful increase in explained variance, then $P^2_{larger} = P^2_{smaller} + \Delta P^2 = .47 + .066 = .536$. Using equation 16.A3.2, we determine that

$$f^2 = \Delta P^2/(1 - P^2_{larger}) = .066/.464 = 0.142.$$

In G*Power, we choose F-tests from the Test family drop-down list and Linear multiple regression: Fixed model, R2 increase from the Statistical test drop-down list. From the Type of power analysis drop-down list, we choose A priori: Compute required sample size - given α, power, and effect size. The effect size, $f^2 = 0.142$, is entered in the Effect size box, and α and power are entered below this. The text box labeled Number of tested predictors asks for the difference in the number of predictors between the larger and smaller models. In our example, this is 2 − 1 = 1. The text box labeled Total number of predictors asks for the number of predictors in the larger (full) model. In our example, this is 2. When the analysis has been described, clicking Calculate returns the result of the analysis in the Output parameters area of the dialog.

Figure 16.A3.2 shows that we require 59 subjects to detect $\Delta P^2 = .066$ with power = .8, when there are two predictors, $P^2_{smaller} = .47$ and α = .05. Of course, these numbers relate to our example of adding *TIE* scores to students' application package. If the researchers had deemed a 6.6% increase in explained variance to be meaningful, they should have had 59 rather than 27 participants in order to achieve power = .8.

The sample size in our *GRE/TIE* example ($n = 27$) suggests that the researchers either (i) were expecting ΔP^2 to be much larger than .066 or (ii) hadn't given much thought to what value of ΔP^2 would be meaningful, and they thus obtained as many subjects as they could and hoped for the best. Planning your sample size based on your best estimate of $P^2_{smaller}$ and having a sense of what ΔP^2 would be meaningful is essential to worthwhile research.