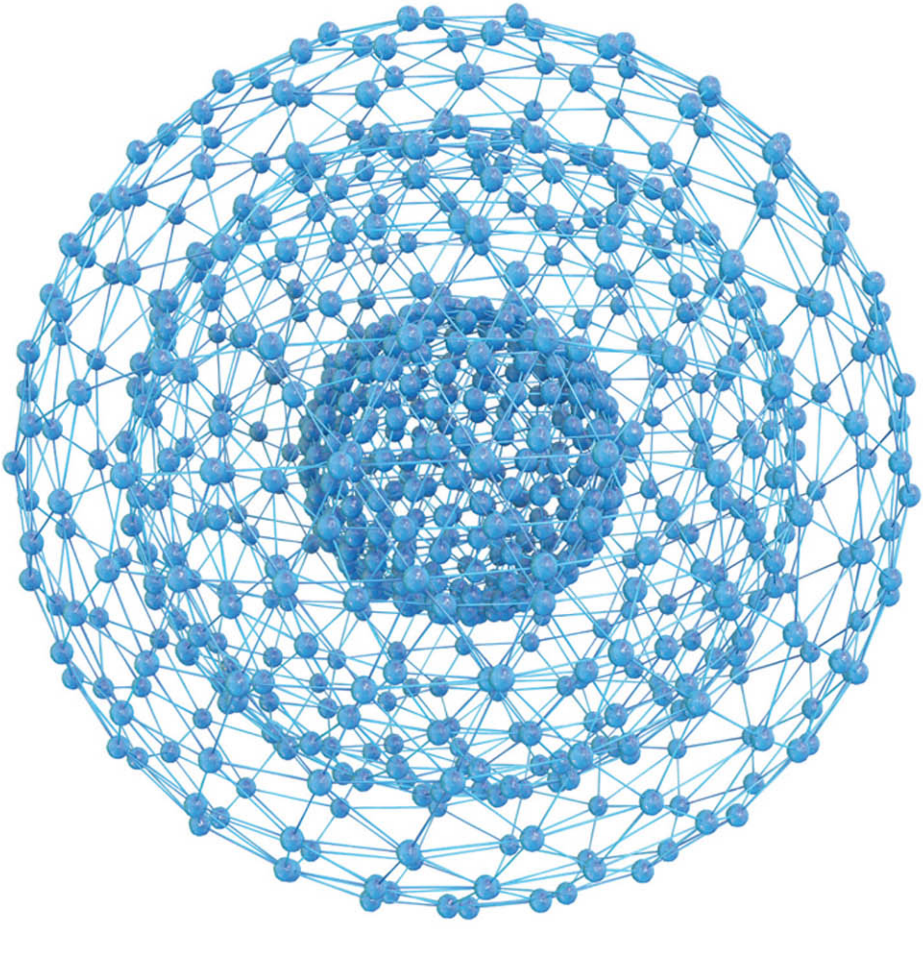
REGRESSION ANALYSIS

Microsoft® Excel®



QUE®

Conrad Carlberg

Contents at a Glance

Regression Analysis Microsoft® Excel®

Conrad Carlberg

QUE®

800 East 96th Street,
Indianapolis, Indiana 46240 USA

Age, you'll account for 32% of the variance in Weight: the simple correlation between Age and Weight is 0.56, and the square of that correlation is 0.32 or 32%.

Having regressed Weight onto Age, you continue by regressing Weight onto Height. Their simple correlation is 0.65, so 0.42 (that's 0.65 squared) or 42% of the variance in Weight is associated with the variance in Height.

By now you have explained 32% of the variance in Weight as associated with Age, and 42% of the variance in Weight as associated with Height. Have you therefore explained 32% + 42% = 74% of the variance in Weight?

No, you probably haven't. The only situation in which that could be true is if Age and Height, your two predictor variables, are *themselves* unrelated and therefore uncorrelated. It's certainly true that two predictor variables can be uncorrelated, but it's almost always because you have designed and conducted an experiment in such a way that the predictor (or explanatory) variables share no variance. You'll see instances of that in later chapters when we take up special coding methods.

In a situation such as the one described in this example, though, it would be pure luck and wildly improbable to find that your predictors, Age and Height, are uncorrelated. You didn't assign your subjects to particular values of Age and Height: That's just how they showed up, with their own ages and heights. Because their correlation with one another is 0.74, Age and Height share 0.74^2, or 55% of their variance.

So when you regress Weight onto Age, you assign some of the variance in Weight to the predictor Age. When you continue the analysis by adding Height to the equation, the variance in Weight shared with Age ought not to be available to share with the predictor Height. To simply add the squared correlations of Age with Weight and Height with Weight would be to count some of the variance twice: the variance that is shared by all three variables.

The solution is to use a semipartial correlation, and thereby to adjust the values of one of the two predictor variables so that they're uncorrelated and share no variance. (Conceptually, anyway—a function such as LINEST() that performs multiple regression is not based on code that follows this sequence of events, but LINEST() emulates them.) If you use a semipartial correlation to remove the effect of one predictor from the other, but leave that effect in place in the predicted variable, you ensure that the variance shared by the predictor and the predicted variable is unique to those variables. You won't double-count any variance.

That last paragraph contains an extremely important point. It helps lay the foundation of much discussion, in later chapters, of how we assess the effects of adding variables to a multiple regression equation. So I repeat it, this time in the context of this section's Height-Age-Weight example.

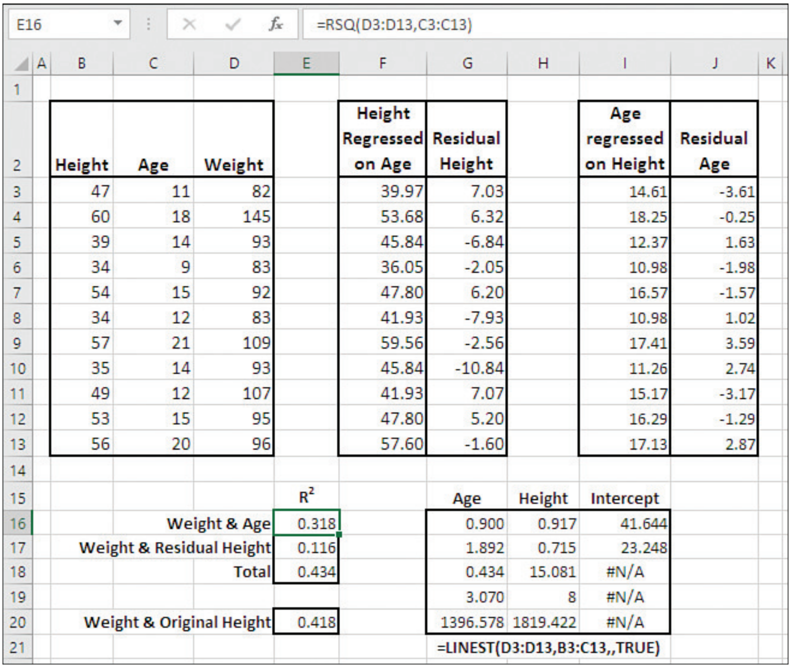
The predictor variables are Height and Age, and the predicted variable is Weight. This issue would not arise if Height and Age were uncorrelated, but they *are* correlated. Therefore, the two predictors share variance with one another. Furthermore, they share variance with the predicted variable Weight—if they didn't, there would be no point to including them in the regression.

The correlation between Height and Weight is 0.65. So if we start out by putting Height into the equation, we account for 0.65^2, or 42% of the Weight variance. Because Height and Age also share variance, some of that 42% is likely shared with Age, along with Height and Weight. In that event, if we just added Age into the mix along with Height, some variance in Weight would be accounted for twice: once due to Height and once due to Age. That would incorrectly inflate the amount of variance in Weight that is explained by the combination of Height and Age.

It can get worse—what if the correlation between Height and Weight were 0.80, and the correlation between Age and Weight were also 0.80? Then the shared variance would be 64% for Height and Weight, and also 64% for Age and Weight. We would wind up explaining 128% of the variance of Weight, a ridiculous outcome.

However, if we apply the notion of semipartial correlations to the problem, we can wind up with *unique* variance, variance that's associated only with a given predictor. We can take the semipartial correlation of Age with Height, partialling Age out of Height (but not out of Weight). See Figure 3.21.

Figure 3.21
We use semipartial correlations to remove the effect of one predictor from the other predictor, but *not* from the predicted variable.



E16 =RSQ(D3:D13,C3:C13)

	A	B	C	D	E	F	G	H	I	J	K
1											
2		Height	Age	Weight		Height Regressed on Age	Residual Height		Age regressed on Height	Residual Age	
3		47	11	82		39.97	7.03		14.61	-3.61	
4		60	18	145		53.68	6.32		18.25	-0.25	
5		39	14	93		45.84	-6.84		12.37	1.63	
6		34	9	83		36.05	-2.05		10.98	-1.98	
7		54	15	92		47.80	6.20		16.57	-1.57	
8		34	12	83		41.93	-7.93		10.98	1.02	
9		57	21	109		59.56	-2.56		17.41	3.59	
10		35	14	93		45.84	-10.84		11.26	2.74	
11		49	12	107		41.93	7.07		15.17	-3.17	
12		53	15	95		47.80	5.20		16.29	-1.29	
13		56	20	96		57.60	-1.60		17.13	2.87	
14											
15					R^2		Age	Height	Intercept		
16				Weight & Age	0.318		0.900	0.917	41.644		
17				Weight & Residual Height	0.116		1.892	0.715	23.248		
18				Total	0.434		0.434	15.081	#N/A		
19							3.070	8	#N/A		
20				Weight & Original Height	0.418		1396.578	1819.422	#N/A		
21							=LINEST(D3:D13,B3:C13,,TRUE)				

Figure 3.21 shows how to use the residuals of one predictor, having partialled out the effects of the *other* predictor. The residuals of Height, after removing from it the effects of Age, appear in the range G3:G13. These cells are of particular interest:

- Cell E16 shows the R^2, the shared variance, between Weight and Age. It's returned easily using Excel's RSQ() function:

  =RSQ(D3:D13,C3:C13)

  At this point we're interested in the R^2 between the actual observations and no variance has been partialled from the first predictor variable (or, for that matter, from the outcome variable Weight).
- Cell E17 shows the R^2 between Weight and the residual values of Height, having already partialled Age out of Height:

  =RSQ(D3:D13,G3:G13)

  We partial Age out of Height so that we can calculate the R^2 between the outcome variable Weight and the *residual* values of Height. We have already accounted for all the variance in Age that's associated with Weight. We don't want to double-count any of that shared variance, so we first partial variance shared by Age and Height out of Height, and then determine the percent of variance shared by Weight and the Height residuals.
- Cell E18 shows the total of the R^2 values for Weight with Age, and for Weight with the residual Height values. They total to 0.434.
- Cell E20 shows the R^2 between Weight and the original Height values. Notice that it is several times larger than the R^2 between Weight and the residual Height values in cell E17. The difference is due to partialling Age out of Height.

Also notice that the total of the R^2 values, in E18, is exactly equal to the value in cell G18. That value, in G18, is the R^2 for the full multiple regression equation, returned by LINEST() and making simultaneous use of Age and Height as predictors of Weight. Had we simply added the raw R^2 values for Age with Weight and Height with Weight, we would have come up with a total R^2 value of 0.318 + 0.418 or 0.736, a serious overestimate.

What if we had started with Height as the first predictor instead of Age? The results appear in Figure 3.22.

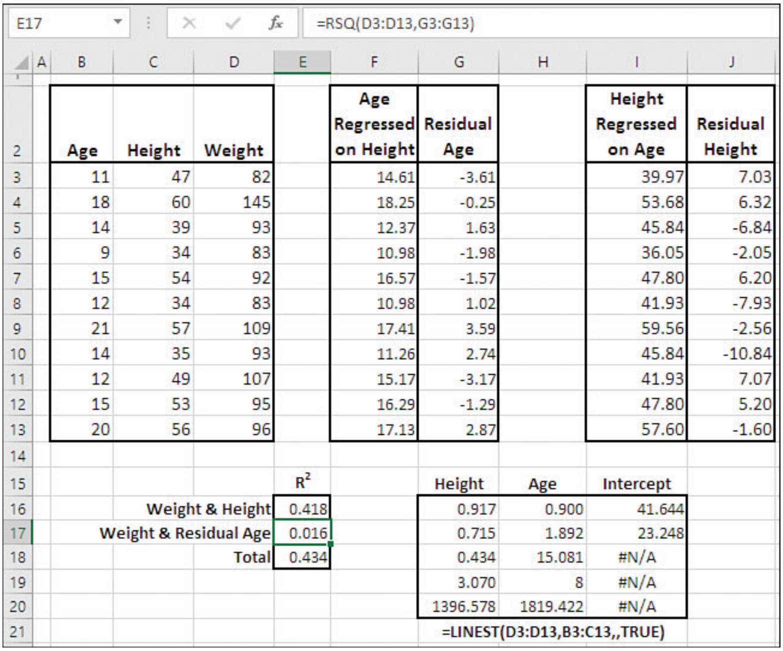
In contrast to Figure 3.21, Figure 3.22 begins by regressing Age onto Height instead of Height onto Age, in the range F3:F13. Then the residual Age values are calculated in G3:G13 by subtracting the predicted Age values from the actual observations in column B.

Then, in Figure 3.22, the unadjusted R^2 for Weight with Height appears in cell E16 (in Figure 3.21, E16 contains the unadjusted R^2 for Weight with Age). Figure 3.22 also supplies in cell E17 the R^2 for Weight with the residual values of Age in G3:G13.

Compare Figures 3.21 and 3.22, and note that the individual R^2 values in G16 and G17 differ. The difference is strictly due to which predictor variable we allowed to retain the variance shared with the other predictor variable: Age in Figure 3.21 and Height in

Figure 3.22. *The total variance explained by the two predictor variables together is the same in both cases.* But the amount of variance in Weight that's attributable to each predictor is a function of which predictor we allow to enter the equation first.

Figure 3.22
The order in which predictors enter the equation affects only the degree of their contribution to the total R^2.



E17 =RSQ(D3:D13,G3:G13)

	A	B	C	D	E	F	G	H	I	J
1										
2		Age	Height	Weight		Age Regressed on Height	Residual Age		Height Regressed on Age	Residual Height
3		11	47	82		14.61	-3.61		39.97	7.03
4		18	60	145		18.25	-0.25		53.68	6.32
5		14	39	93		12.37	1.63		45.84	-6.84
6		9	34	83		10.98	-1.98		36.05	-2.05
7		15	54	92		16.57	-1.57		47.80	6.20
8		12	34	83		10.98	1.02		41.93	-7.93
9		21	57	109		17.41	3.59		59.56	-2.56
10		14	35	93		11.26	2.74		45.84	-10.84
11		12	49	107		15.17	-3.17		41.93	7.07
12		15	53	95		16.29	-1.29		47.80	5.20
13		20	56	96		17.13	2.87		57.60	-1.60
14										
15					R^2		Height	Age	Intercept	
16				Weight & Height	0.418		0.917	0.900	41.644	
17				Weight & Residual Age	0.016		0.715	1.892	23.248	
18				Total	0.434		0.434	15.081	#N/A	
19							3.070	8	#N/A	
20							1396.578	1819.422	#N/A	
21							=LINEST(D3:D13,B3:C13,,TRUE)			

At this point that might seem a trivial issue. What's important is how accurately the overall regression equation performs. The contribution of individual variables to the total explained variance is by comparison a relatively minor issue.

Except that it's not. When you begin to consider whether to even use a variable in a multiple regression equation, it's a relatively major issue. It can affect your assessment of whether you've chosen the right model for your analysis. I'll take those matters up in some detail in Chapter 5. First, though, it's necessary to add Excel's LINEST() function to this book's toolkit. The LINEST() function is critically important to regression analysis in Excel, and Chapter 4 discusses it in much greater detail than I have thus far.

This page intentionally left blank

Using the LINEST( ) Function

4

IN THIS CHAPTER

The worksheet function LINEST() is the heart of regression analysis in Excel. You could cobble together a regression analysis using Excel's matrix analysis functions without resorting to LINEST(), but you would be working without your best tools. So here you are, three chapters into a book entirely about regression, and I haven't even said anything about how to enter Excel's most important regression function into a worksheet.

I'll get into that next, but first let me explain an apparent contradiction. Projected into its smallest compass, with just one predictor variable, LINEST() returns a maximum of 10 statistics. Yet you can use LINEST() to perform something as simple as a test of the difference between the means of two groups, or as complex as a factorial analysis of covariance, complete with factor-by-factor and factor-by-covariate interactions. You can use it to run what's termed *model comparison*, which enables you to assess the statistical effect of adding a new variable (or of removing an existing variable) from a regression equation. You can use LINEST() to perform curvilinear regression analysis and orthogonal contrasts.

It's all in the way that you arrange the raw data before you point LINEST() at it. I'll get into that issue in subsequent chapters. First, though, it's important to understand the mechanics of putting LINEST() on the worksheet, as well as what those ten statistics mean and how they interact with one another. That's the purpose of this chapter.

Let's start with the mechanics.

Array-Entering LINEST( )

Different functions in Excel react in different ways when you array-enter them, by way of Ctrl-Shift-Enter, compared to how they