Chapter 7:

Simple curvilinear models

Advance organizer

In previous chapters we have talked about the assumption in almost any type of linear modeling that the model is, well, *linear*. Whether we are talking about OLS regression or logit or probit models, linear is in the label. However, I think that in many areas of science, this assumption may not be tenable or even desirable. I believe that if we routinely looked for curvilinear relationships, we would find many. In fact, while writing this chapter, I had to explore surprisingly few examples to produce the curvilinear results shown herein. The fact of the matter is that curves are everywhere, and I hope this chapter encourages you to begin looking for them in your data. You will find that it is not terribly painful, and can produce much more nuanced and interesting results.

In this chapter we will briefly review the concept of curvilinearity, how to test for curvilinearity more formally, how to account for curvilinearity in your regression analyses, and how to graph curvilinear effects.[1] At the end of the chapter I will also digress into a brief section on how to have even more fun with curvilinear effects if you know a little calculus (I think everyone should, by the way).

In this chapter we will cover:

- Basic concepts in curvilinear effects
- Curvilinear effects in OLS regression
- Curvilinear effects in logistic regression
- Examples of APA-compliant summaries of analyses
- Guidance on how to perform these analyses in various statistical packages will be available on-line at http://jwosborne.com.

Zeno's paradox, a Nerdy science joke, and inherent curvilinearity in the universe…

My high school science teacher, Larry Josbeno was not only a brilliant teacher, but was also fond of lousy physics jokes. One of his favorites related to Zeno's paradoxes and was a variant of what is apparently a classic mathematical joke, which I paraphrase below (but cannot replicate his spot-on delivery):[2]

> *At a high school dance, a group of boys are lined up on one wall of a dance hall, and an equal number of girls are lined up on the opposite wall 10 meters apart. Both groups are then instructed to advance toward each other by one half the distance separating them every ten seconds (i.e., if they are distance d apart at time 0, they are d/2 at time=10, d / 4 at time=20, d /8 at time = 30, and so forth.) A mathematician, a physicist, and an engineer are asked when they would meet at the center of the dance hall. The mathematician said they would never actually meet because the series is infinite. The physicist said they would meet when time equals infinity. The engineer said that within one minute they would be close enough for all "practical" purposes.*

[1] I believe graphical representations of complex findings like curvilinear effects and interaction effects (where found) are critical to effectively communicating the results of research to the audience of interest.

[2] *(see Field, Paul and Weisstein, Eric W. "Zeno's Paradoxes." From MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/ZenosParadoxes.html)*

Enthusiastic adolescent laughter ensued, predictably. Thank you, Mr. Josbeno! But what does this have to do with curvilinear effects in regression? Like many things in life, if we were to explore the relationship between time and distance between our girls and boys, the relationship is not linear, as Figure 7.1 below shows. And curvilinearity is the topic of this chapter!

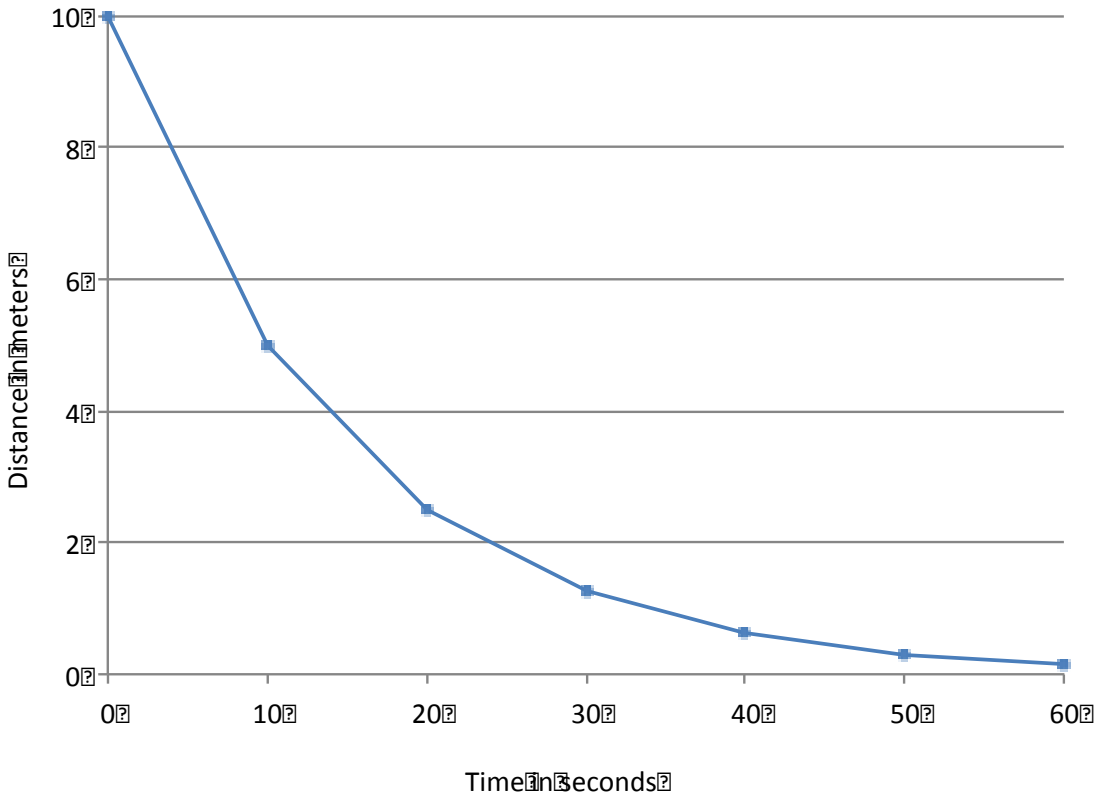

Figure 7.1 Zeno's paradox in the high school dance

A brief review of simple algebra

Curves in algebra are relatively simple to understand. There are an infinite number of specialized types of curves, but we can advance the field of statistical methods greatly by beginning to apply some simple principles. We have seen many examples of lines where we have a Y and an X, such as in Equation 7.1, where we have an intercept (when X = 0, Y = 15) and a slope (for every increment of X, Y will decrease by 8):

EQ. 7.1

$$Y = 15 - 8X$$

When you don't see a superscript or power next to a variable, that variable is assumed to be raised to the first power (e.g., X^1). Any variable raised to the first power is the same as that variable (i.e., $X^1 = X$), just as multiplying any variable by 1 is the same value (1X = X). This is why 1 is called the multiplicative identity in mathematics.[3] You can see the line graphed in Figure 7.2, looking remarkably like many regression graphs we have seen.

[3] 0 is the additive identity- X+0 =X. Raising anything to the 0 power makes it equal to 1 (i.e., $X^0 = 1$).

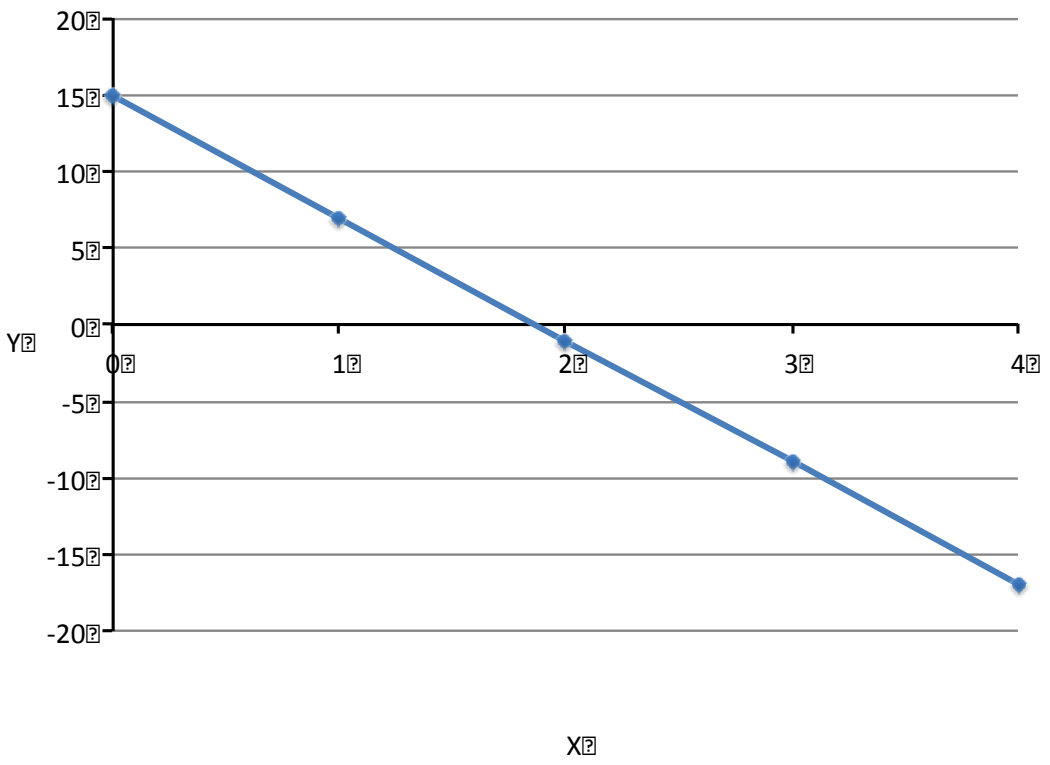

Figure 7.2. Graph of Equation 7.1.

When we have a variable raised to a power, we get a curve. For example, if we have a squared term in the equation (this then becomes a quadratic equation), as in Equation 7.2, we get a curve with one inflection point (or change in direction), which is graphed in Figure 7.3.

EQ. 7.2

$$Y = 15 - 8X + 2X^2$$

As you can see in Figure 7.3, there is one point, at X = 2, where the slope of the curve is 0 (flat), the point where the slope changes from negative to positive. That is the only inflection point for the curve. If you graph it infinitely in each direction the curve will have no other.

If we examine an equation with a cubic term (a variable raised to the third power), we then get two inflection points. For example, Equation 7.3, graphed in Figure 7.4, has X raised to the third power, and will therefore have two inflection points.

EQ. 7.3

$$Y = X + X^2 - 0.5X^3$$

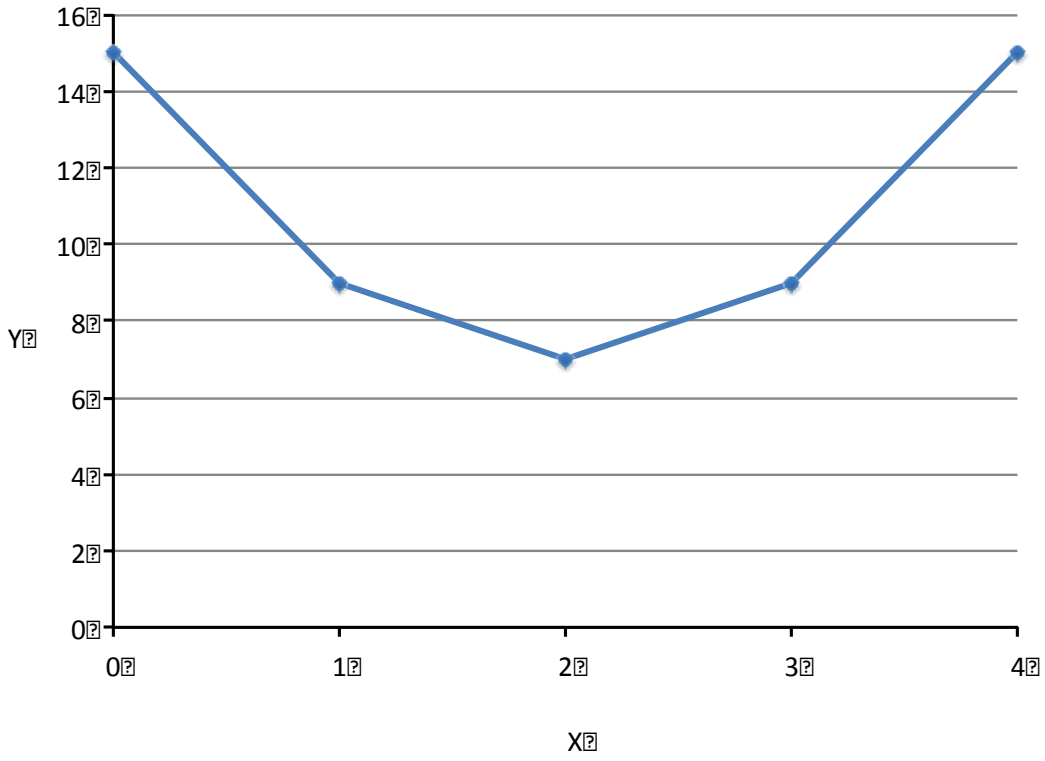

Figure 7.3. Graph of Equation 7.2.

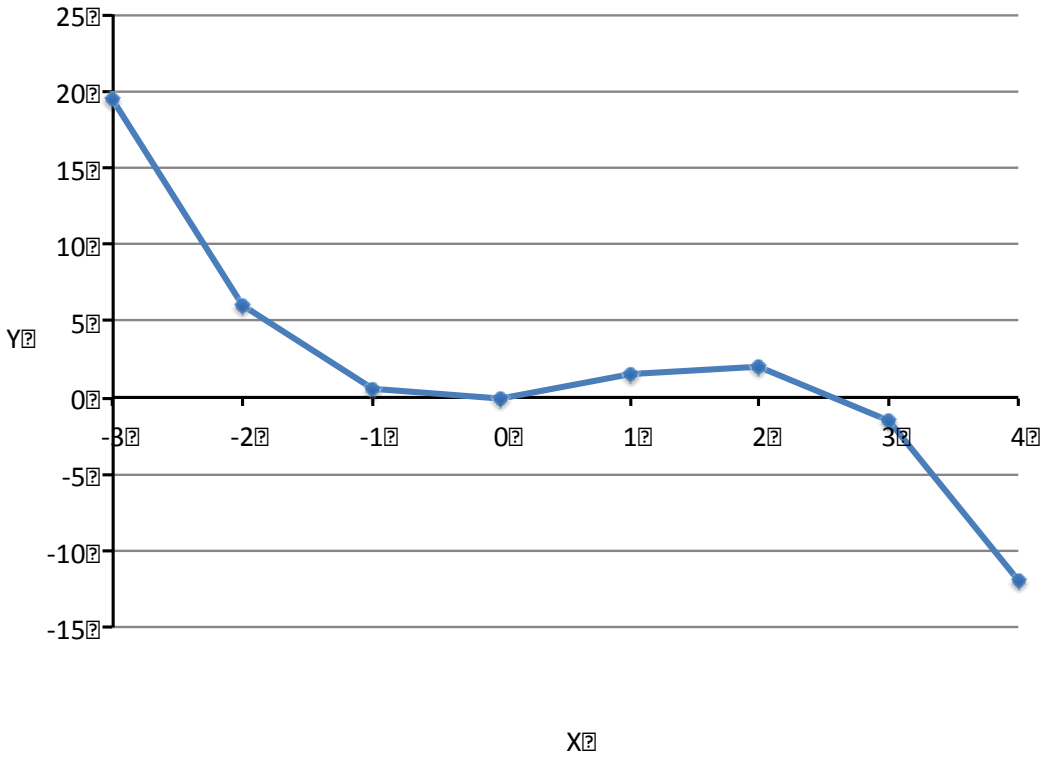

Figure 7.4. Graph of equation 7.3

At the end of the chapter I will review some more fun you can have with curves using calculus, but for now the general principle is that an equation with a variable raised to a particular power (k) will produce a curve with $k - 1$ inflection points. Keep in mind that lines extend infinitely in both directions, and the inflection points you expect may not occur within the observed range.

One example in this chapter will return to the AAUP data from earlier chapters and the relationship between size of a university (total number of faculty, NUM_TOT, z-scored) and associate professor salary (SAL_AP, measured in hundreds of dollars). As we saw in Chapters 2 and 3, these data are problematic due to non-normality of residuals, heterogeneity, etc. This can sometimes be caused models that fail to include curvilinearity when it is present in the data.

In Figure 7.5, we see the plot of ZPRED vs. ZRESID (standardized predicted values on the X axis and standardized residuals on the Y axis) from the simple OLS regression predicting SAL_AP from NUM_TOT. Graphs like these can be used to evaluate linearity and homogeneity assumptions, and this one seems to suggest that there might be significant unmodeled curvilinearity in the data.

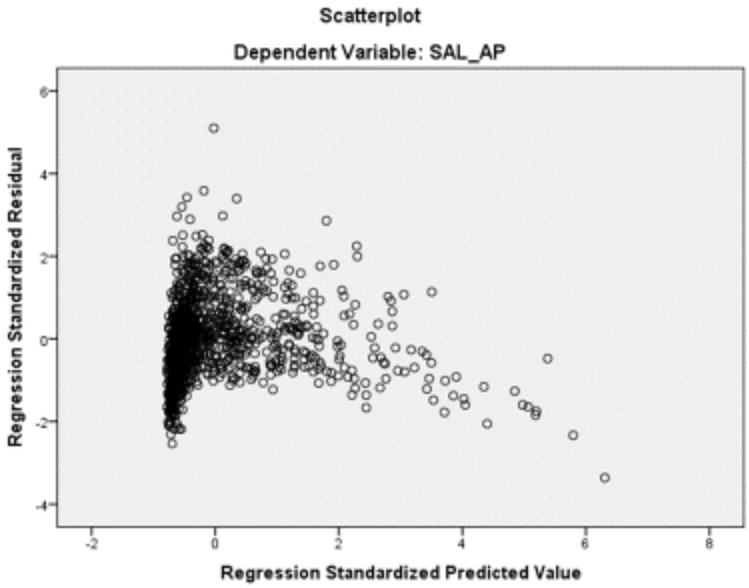

Figure 7.5: Plot of ZPRED vs. ZRESID from the AAUP data analysis predicting salary of associate professors from size of the institution.

We will explore another example within logistic regression will be the curvilinear effect of age (AGE) on the probability of being diagnosed with diabetes from NHIS 2010 (http://www.cdc.gov/nchs/nhis/nhis_2010_data_release.htm).

Hypotheses to be tested

In Equation 7.4, we have an example of a regression equation with a quadratic term in it. In this example, b_2 is the portion of the effect of X_1 on Y that is quadratic, b_3 would be the cubic aspect of the effect, and so on :

EQ. 7.4:

$$\hat{Y} = b_0 + b_1X_1 + b_2X_1^2 + b_3X_1^3 + \ldots + b_kX_1^k$$

If we fail to test for curvilinearity, we are implicitly asserting that b_2 to b_k are equal to 0. Sometimes they are—there are many linear relationships in the world. However, it is poor practice to simply assert an effect without testing for it. Thus, when we engage in curvilinear analyses, we explicitly test hypotheses about those aspects of the effect that are nonlinear. For example, we can test:

H_0: $b_2 = 0$
H_a: $b_2 \neq 0$

H_0: $b_3 = 0$
H_a: $b_3 \neq 0$

and so forth. I generally explore curves in the quadratic and cubic range, and not higher, primarily because I suspect that our data in the social and behavioral sciences are not precise enough to confidently model (and

replicate) more complex curves. I also begin to have trouble explaining more complex curves rationally. But there is no reason why they cannot be modeled if you can explain them and can replicate them

If we reject any of these null hypotheses then we are declaring our assumption of linearity voided, and we should model the curvilinearity rather than ignoring it.

Illegitimate causes of curvilinearity

In this chapter I am most concerned with modeling legitimately curvilinear relationships. As I briefly mentioned in Chapter 02 on assumptions, there are several potential sources of curvilinearity that are not, in my mind, legitimate: model misspecification (omission of important variables), converting interval or ratio variables to ordinal variables with unequal intervals, and uncleaned data (i.e., containing influential data points).

Model misspecification: Omission of important variables. When discussing the assumption that we have correctly specified the model, we introduced the assumption that we have included all relevant and important variables in the model, and have not included extraneous variables. It is possible that omission of important variables can lead to either of these situations. Thus, theory and prior research should help guide you in designing research that accounts for important variables (i.e., prior academic experiences in studying education, prior health events in studying current health status).

Poor data cleaning. As also mentioned previously, I have occasionally seen curvilinear effects arise (or masked) merely because of poor data cleaning—a prominent outlier in one range of the data where there are few other cases can pull the regression line in that area out of linearity leading to the appearance of a curvilinear effect when in fact it is merely poor data cleaning. I suppose it is also possible for highly non-normal data to have the appearance of curvilinearity.

Thus, I would argue that prior to examining data for curvilinear effects, one should be sure that the appropriate variables are modeled in the equation, and that you have done due diligence in data cleaning. Once you have satisfied those basic steps (which should probably be part of any analysis regardless of whether curvilinearity is suspected), it is time to explore whether curvilinear effects exist in the data. There are examples of how data cleaning can reveal an existing curvilinear effect at the end of the chapter.

Detection of nonlinear effects.

Theory. First and foremost, theory and common sense is always a good guide. I tend to believe that many things in social science (and health sciences as well) are curvilinear in nature, and so I routinely check for these effects. If prior research has indicated curvilinear effects, or if there is good cause to suspect that the effect might not be uniform across the entire range of a variable, it is probably worth taking a few minutes to test.

Ad hoc testing. They are easily tested by entering X, X^2, and X^3 terms into an equation. In my experience, if there is curvilinearity, adding squared and cubed terms tends to capture much of the curvilinearity if there is any.

Box-Tidwell transformations. Those preferring a more strategic approach to this issue may enjoy exploring Box-Tidwell transformations (Box & Tidwell, 1962) as a more methodical approach to testing and specifying curvilinear effects (and more importantly, linearizing relationships). Many prominent regression authors and texts (i.e., Cohen, Cohen, West, & Aiken, 2002, pp. 239-240) suggest Box-Tidwell as a method of easily exploring whether any variables have non-linear effects. However, I have rarely seen studies published in journals or books use it. I personally only used it for the preparation of this book. That is not to say it should not be used,

The essential process for Box-Tidwell is to (a) perform an initial analysis with the independent variables of interest in the regression equation, (b) transform all independent variables of interest via Box-Tidwell, below, (c) enter them into the regression equation simultaneously along with the original untransformed variables, and (d) see which of the transformed variables (if any) are significant.

The Box-Tidwell transformation is expressed in Equation 7.5:

EQ. 7.5:

$$V_i = X(\ln X).$$

After computing variable V, it is entered into the equation after X is in the equation, as in Equation 7.6a (OLS regression) or Equation 7.6b (logistic regression):

EQ 7.6a:

$$\hat{Y} = b_0 + b_1X_1 + b_2V_1$$

EQ 7.6b:

$$\text{logit}(\hat{Y}) = b_0 + b_1X_1 + b_2V_1$$

The null hypothesis to be tested here is that there is no significant curvilinearity in the data:

H_0: $b_2 = 0$
H_a: $b_2 \neq 0$

If we reject the null hypothesis, we must recognize that there is curvilinearity in the data. Further, Box-Tidwell provides for a way to estimate the nature of the curvilinear effect, as Equation 7.7 shows:

EQ. 7.7

$$\hat{\lambda} = \frac{b_2}{b_1} + 1$$

Where b_2 is taken from the second analysis and b_1 is taken from the initial analysis without the V in the equation. You can do successive iterations of this process as well, entering $X^{\hat{\lambda}}$ in place of the original X in both the original steps and the calculation of V but in my opinion that tends to over fit the model unnecessarily and undermine replicability. Our data in the social sciences are not the same character and nature as in the physical sciences and manufacturing, for example.

Basic principles of curvilinear regression

One issue we will run into with curvilinear regression and also in the future when we examine interactions, is that X^2 and X^3 are *collinear* with X. In other words, they are highly correlated with the original variable. Thus, we cannot easily evaluate lower-order variables when higher-order variables are in the equation as high collinearity can distort estimates. We will discuss collinearity in more detail in the next chapter. For now we will introduce a couple important practices to allow you to effectively explore curvilinearity in your data.

Occam's razor. One general principle in statistical science that we usually work from is that we want to find simultaneously the best explanation for a phenomenon and the simplest. Thus, when one variable is sufficient to explain a phenomenon, we do not include ten. Likewise, if a linear equation can effectively explain a relationship, we prefer that to a curvilinear relationship, and we prefer a quadratic to cubic (or other exotic curve) so long as the models are largely equivalent in strength.

Ordered entry of variables. Let us say that we have X, X^2, and X^3. Given our preference for simpler variables over more complex variables, we will enter X into the analysis first, and then enter X^2. If X^2 adds a significant increment to the model, we will keep it. If not, we will not retain it in the model. If X^2 is significant and we enter X^3, it should significantly improve the model over the simpler effects or not be retained.

Each effect is one part of the entire effect. Remember that X, X^2, and X^3 are all different aspects of the effect of the single variable X. It is an artifact of how we create linear regression equations that we have to model different aspects of the variable in this way. It is important to remember that X^2 is just one aspect of X. This is partly the reason for the ordered entry that we just discussed. We can only evaluate X^2 when X is already in the equation because that allows us to separate out the linear and quadratic effects. We can only

evaluate the effect of X^3 when X and X^2 are already in the equation. Conversely, we cannot evaluate the effect of X when X^2 or X^3 are in the equation because of collinearity.

Centering. We have briefly discussed the benefits of centering (I usually do it through conversion to *z* –scores but you can do it other ways also and achieve the same benefit) in prior chapters. In this and future chapters, centering will become mandatory. Aiken and West, in their seminal book on regression (Aiken & West, 1991; Cohen et al., 2002) defined centering as critical for any analysis where curvilinear effects (or interactions, as we will discuss in Chapter 9) are found. There are some technical reasons for this. I will refer you to their excellent books if you are interested in that level of detail. The bottom line is that we will center X before calculating X^2 or X^3 or any other curvilinear transform of X.

Curvilinear OLS regression example: Size of the University and faculty salary

Our first example returns to our example of "problematic" data from Chapter 2: the AAUP data on institutional size (zNUM_TOT) and the salary of associate professors (SAL_AP; salary in hundreds of dollars). Because the residual plot in Figure 7.5 looked so weird, I computed quadratic, cubic, and quartic components of zNUM_TOT ($zNUM\_TOT^2$, $zNUM\_TOT^3$, $zNUM\_TOT^4$) to demonstrate the basic principles of testing for curvilinear effects in OLS regression. Each term was entered on its own step, so that there are four versions of the regression line equation, as you can see in Equations 7.8a-d:

EQ. 7.8a:

$$SAL\_AP = b_0 + b_1 zNUM\_TOT + e$$
$$H_0: \Delta R^2 = 0; b_1 = 0$$
$$H_a: \Delta R^2 \neq 0; b_1 \neq 0$$

EQ. 7.8b:

$$SAL\_AP = b_0 + b_1 zNUM\_TOT + b_2 zNUM\_TOT^2 + e$$
$$H_0: \Delta R^2 = 0; b_2 = 0$$
$$H_a: \Delta R^2 \neq 0; b_2 \neq 0$$

EQ. 7.8c:

$$SAL\_AP = b_0 + b_1 zNUM\_TOT + b_2 zNUM\_TOT^2 + b_3 zNUM\_TOT^3 + e$$
$$H_0: \Delta R^2 = 0; b_3 = 0$$
$$H_a: \Delta R^2 \neq 0; b_3 \neq 0$$

EQ. 7.8d:

$$SAL\_AP = b_0 + b_1 zNUM\_TOT + b_2 zNUM\_TOT^2 + b_3 zNUM\_TOT^3 + b_4 zNUM\_TOT^4 + e$$
$$H_0: \Delta R^2 = 0; b_4 = 0$$
$$H_a: \Delta R^2 \neq 0; b_4 \neq 0$$

As a point of methodological information, this type of user-controlled entry where we enter one term at a time as specified by the statistician is traditionally called "hierarchical entry." We will discuss more about different methods of entry in Chapter 8.

The results from this analyses are presented in Table 7.1 in a more expanded form than normal so that we can explore all the nuances of this analysis. At each step, as a new term is entered into the model, we are testing the null hypothesis that the model is not improved, and that the new term has no relationship to the outcome variable. The alternative hypothesis is that the model is significantly improved, and that the new term is significantly related to the outcome.

Table 7.1
Curvilinear analysis of AAUP data

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	ZNUM_TOT	.	Enter
2	zNUM_TOT^2	.	Enter
3	zNUM_TOT^3	.	Enter
4	zNUM_TOT^4	.	Enter

a. Dependent Variable: SAL_AP

Model Summary[e]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.494[a]	.244	.244	62.223	.244	362.895	1	1123	.000
2	.586[b]	.344	.342	58.018	.099	169.688	1	1122	.000
3	.618[c]	.382	.380	56.318	.038	69.763	1	1121	.000
4	.643[d]	.413	.411	54.896	.031	59.829	1	1120	.000

a. Predictors: (Constant), ZNUM_TOT
b. Predictors: (Constant), ZNUM_TOT, zNUM_TOT^2
c. Predictors: (Constant), ZNUM_TOT, zNUM_TOT^2, zNUM_TOT^3
d. Predictors: (Constant), ZNUM_TOT, zNUM_TOT^2, zNUM_TOT^3, zNUM_TOT^4
e. Dependent Variable: SAL_AP

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1405040.683	1	1405040.683	362.895	.000[b]
	Residual	4347982.294	1123	3871.756		
	Total	5753022.978	1124			
2	Regression	1976232.584	2	988116.292	293.547	.000[c]
	Residual	3776790.394	1122	3366.123		
	Total	5753022.978	1124			
3	Regression	2197501.684	3	732500.561	230.946	.000[d]
	Residual	3555521.294	1121	3171.741		
	Total	5753022.978	1124			
4	Regression	2377801.174	4	594450.293	197.256	.000[e]
	Residual	3375221.804	1120	3013.591		
	Total	5753022.978	1124			

a. Dependent Variable: SAL_AP
b. Predictors: (Constant), ZNUM_TOT
c. Predictors: (Constant), ZNUM_TOT, zNUM_TOT^2
d. Predictors: (Constant), ZNUM_TOT, zNUM_TOT^2, zNUM_TOT^3
e. Predictors: (Constant), ZNUM_TOT, zNUM_TOT^2, zNUM_TOT^3, zNUM_TOT^4

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	415.641	1.856		223.997	.000	412.001	419.282
	ZNUM_TOT	35.065	1.841	.494	19.050	.000	31.453	38.676
2	(Constant)	427.239	1.946		219.571	.000	423.421	431.057
	ZNUM_TOT	66.442	2.958	.936	22.465	.000	60.638	72.245
	zNUM_TOT^2	-12.076	.927	-.543	-13.026	.000	-13.895	-10.257

3	(Constant)	439.712	2.408		182.617	.000	434.988	444.437
	ZNUM_TOT	81.497	3.390	1.149	24.041	.000	74.846	88.149
	zNUM_TOT^2	-35.339	2.927	-1.589	-12.074	.000	-41.081	-29.596
	zNUM_TOT^3	4.075	.488	.903	8.352	.000	3.117	5.032
4	(Constant)	451.091	2.770		162.852	.000	445.656	456.526
	ZNUM_TOT	79.348	3.316	1.118	23.929	.000	72.842	85.855
	zNUM_TOT^2	-69.306	5.237	-3.116	-13.234	.000	-79.581	-59.031
	zNUM_TOT^3	21.317	2.279	4.726	9.352	.000	16.845	25.789
	zNUM_TOT^4	-1.996	.258	-2.367	-7.735	.000	-2.502	-1.490

a. Dependent Variable: SAL_AP

As you can see in the first part of Table 7.1, we do indeed have four different models being tested, each one adding a higher-powered version of the z-scored zNUM_TOT. Having requested the change statistics in the model summary, we can see whether the model significantly improves with the addition of each new term. You can see in the first line of the model summary that the linear relationship between institution size and faculty salary is relatively strong, accounting for about one-quarter of the variance in salary ($R^2 = 0.244$), and that the model significantly improved (over an empty or intercept-only model; $F_{(1, 1123)} = 362.90$, $p < 0.0001$). For this equation, we can reject the null hypothesis for the overall model. Likewise, the model improved significantly with the addition of the second term, zNUM_TOT^2 ($\Delta R^2 = 0.099$, $F_{(1, 1122)} = 169.69$, $p < 0.0001$), third term, zNUM_TOT^3 ($\Delta R^2 = 0.038$, $F_{(1, 1121)} = 69.76$, $p < 0.0001$), and also the fourth term, zNUM_TOT^4 ($\Delta R^2 = 0.031$, $F_{(1, 1120)} = 59.83$, $p < 0.0001$). Thus, each term adds to a stronger model. This is not normal in my experience, but is welcome as an example for this chapter.

The ANOVA table summarizes the significance of the overall model (testing the null hypothesis that all coefficients are simultaneously zero). This is not usually of interest but you can see the residual sum of squares (unexplained variance) decreasing with each step, meaning that each additional term added to the model is reducing the unexplained (error) variance, and that all models are significant overall. This is generally a good thing. Was the overall model to be non-significant, we could not reject the null hypothesis that the regression coefficients are zero, and could not legitimately discuss any effects.

Finally, the last part of the table summarizes the details of each effect, which is what we are usually interested in. I mentioned earlier in the chapter that when exploring complex effects like curvilinear terms, we generally only interpret the term entered on each step. Because we left the DV in its original metric (hundreds of dollars) we can interpret this as the average salary of an associate professor at the intercept (salary at an average-sized university) is $41,564.10.[4] The linear effect is significant (which we already saw from the model summary table), and has an unstandardized regression coefficient of 35.07 (with a 95%CI of [31.45, 38.68]). We can interpret this as the average salary of an associate professor increases $3,507.00 (and we are 95% sure the true increase in the population is between $3,145.30 and $3,867.60 based on our 95%CI) for each increase of one standard deviation in size. However, this effect is modified by the subsequent significant curvilinear effects entered on the next steps.

Once the quadratic term is entered, we know that the model improves significantly (accounting for almost 10% more variance), and now we have the regression equation details based on the intercept and the two regression coefficients. Note that the intercept has changed, as has the effect of zNUM_TOT, because now that variable is interpreted as the linear effect controlling for the curvilinear effect. These variables are also highly collinear (correlated over $r = 0.81$), and thus strongly effect each other when both are in the equation.

Of course, this quadratic effect is modified by the cubic effect, which is also significant, and so on. In the service of brevity, I usually report this full table, but only interpret the last equation with a significant effect (in this case, the fourth equation, as that effect was also significant).

Note that we always seek to meet assumptions of whatever analysis we are performing. In Figure 7.6, you can see that there are standardized residuals that extend beyond ± 3.0, but the normality is not severely violated (skew = 0.69, kurtosis = 0.92). The scatterplot of the standardized predicted values against the standardized residuals in Figure 7.7 shows substantial improvement from the previous linear regression

[4] Keep in mind these data are many years out of date. Most professors these days make several dollars more per year than back in the 1990s.

analysis (Figure 7.5). Prior to adding the curvilinear terms to the equation, the plot indicated potential nonlinearity and heterogeneity. While this is not perfect, it is much improved, particularly in light of the fact that we have not performed data cleaning yet.

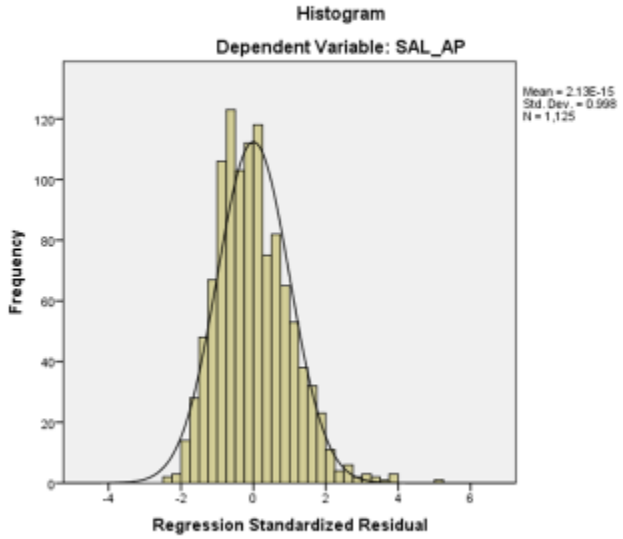

Figure 7.6: Histogram of the standardized residuals from the AAUP curvilinear regression analysis predicting salary of associate professors from size of the institution.

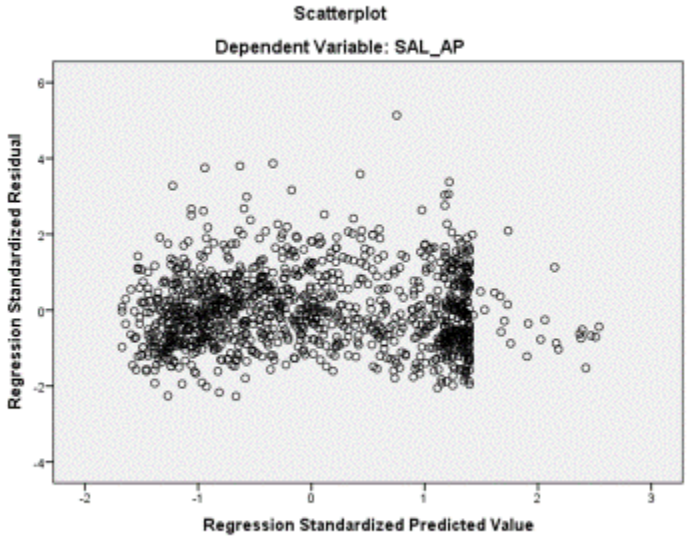

Figure 7.7: Plot of ZPRED vs. ZRESID from the AAUP curvilinear regression analysis predicting salary of associate professors from size of the institution.

Data cleaning

After all terms were in the equation, I examined both standardized residuals and DfFit statistics. DfBetas would also be appropriate to examine, as would other indicators of influence. There was one case that had an exceptionally large DfFit (it was over 33 standard deviations from the mean DfFit value), and that case was removed. The standardized residuals ranged from -2.27 to 5.14, with ten cases (out of 1125) having standardized residuals greater than 3.0. These eleven cases were also removed, resulting in less than 1% of the sample being removed. The results after cleaning are presented (in less detail) below in Table 7.2 and Figures 7.8 and 7.9. The normality of the residuals has improved substantially (skew has now been reduced to 0.38 and the kurtosis is now -0.25). Further, the scatterplot (Figure 7.9) is slightly improved over Figure 7.8, more clearly meeting the assumption of homoscedasticity.

Of course, we would not expect removal of less than one percent of the sample to have massive effects on the results, but we do see incremental improvement in the model. Perhaps more importantly, we can be somewhat confident that the curvilinear effects are truly in the population, and are not the result of a few highly influential cases. You can see that the models are all still significant, meaning that the basic nature of the curve has not changed following data cleaning. Second, you can see that the overall model accounts for slightly more variance than previously.[5] You can also see that the regression coefficients were modified slightly, which is also to be expected if we removed inappropriately influential cases.

Interpreting curvilinear effects effectively

Looking at the coefficients, is it immediately apparent what the curve will look like? I used to be pretty good at algebra and calculus, but my recommendation to you is to graph complex effects like curves. I recommend that you graph a reasonable range, and use it to help you (and more importantly, your reader) intuitively interpret the effect. I provide a graph of all four equations from our analysis below in Figure 7.10 only so you can see the differences between the curves, and the value of taking a few minutes of your time to explore curvilinearity. You would normally only graph the final curve and explain that to your reader. I am also graphing the curve across an expanded range (-4 to +4 SD) so that you can see more of the curve. However, graphing an effect beyond the range of the data is difficult to justify and to replicate. I would recommend keeping it between -3 and +3 at most.

[5] This is because we removed the 11 cases out of over 1100 that were contributing the most error variance.

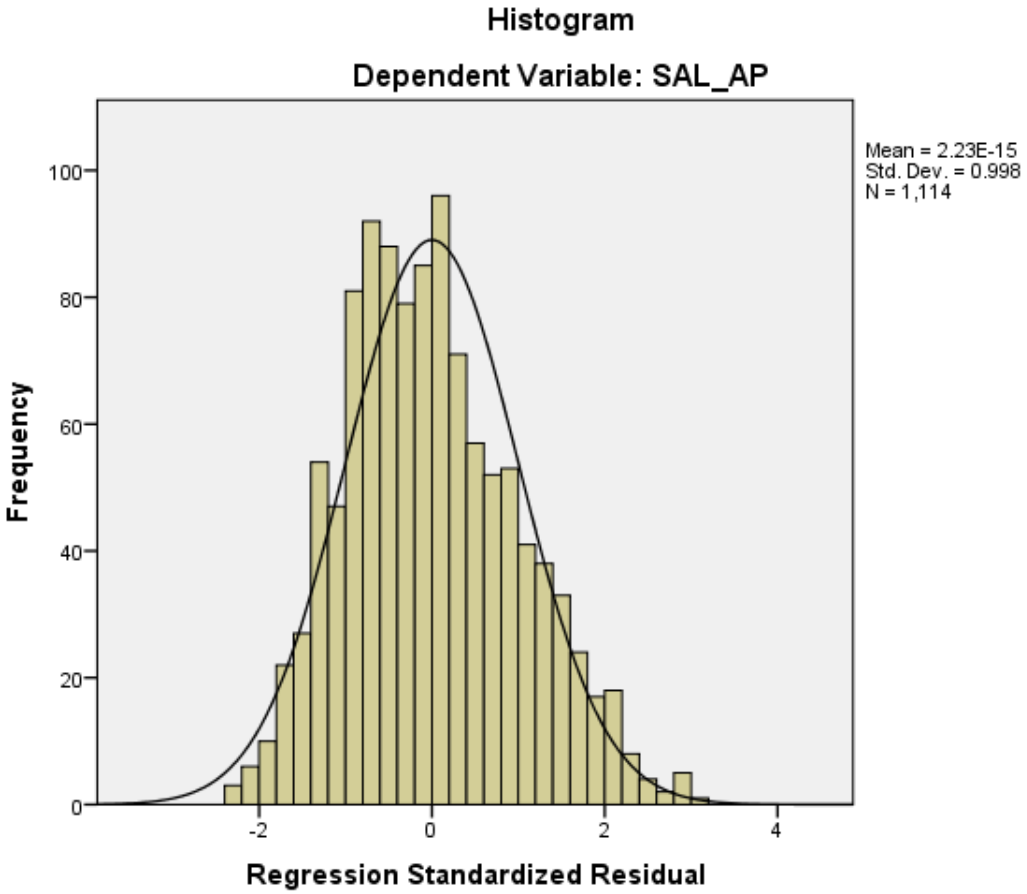

Figure 7.8: Histogram of standardized residuals following data cleaning.

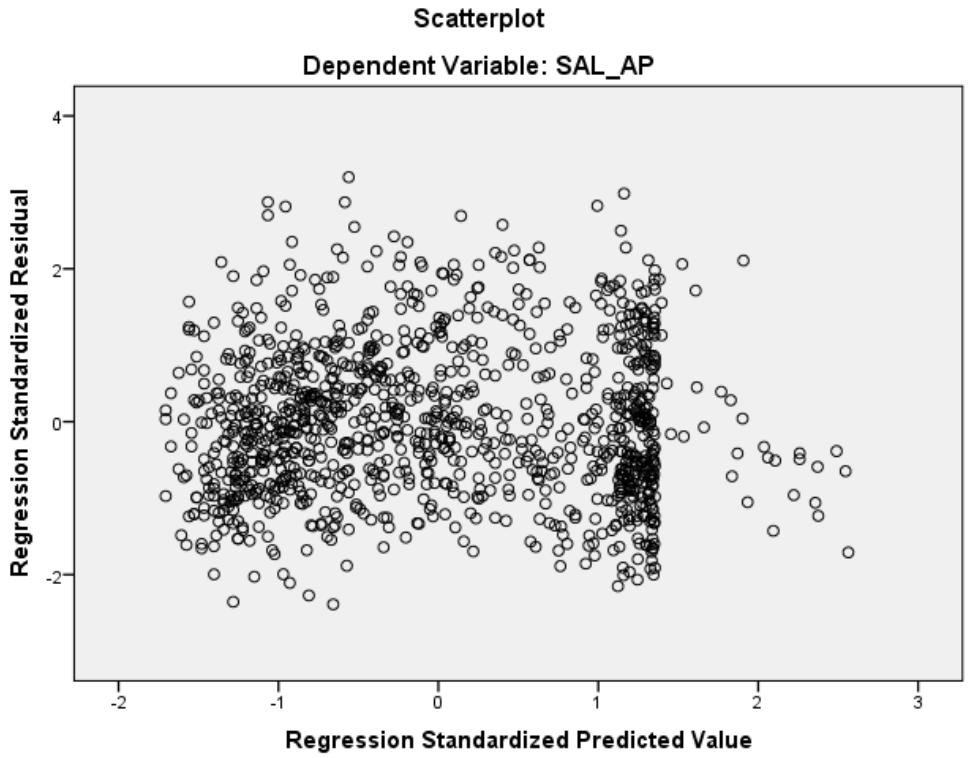

Figure 7.9: Scatterplot of standardized residuals and standardized predicted values after data cleaning.

Table 7.2
Curvilinear analysis of AAUP data after data cleaning

Model Summary[e]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.518[a]	.268	.268	59.075	.268	408.075	1	1112	.000
2	.605[b]	.366	.364	55.041	.097	169.940	1	1111	.000
3	.643[c]	.413	.412	52.958	.048	90.114	1	1110	.000
4	.665[d]	.442	.440	51.648	.029	58.029	1	1109	.000

a. Predictors: (Constant), ZNUM_TOT
b. Predictors: (Constant), ZNUM_TOT, zNUM_TOT^2
c. Predictors: (Constant), ZNUM_TOT, zNUM_TOT^2, zNUM_TOT^3
d. Predictors: (Constant), ZNUM_TOT, zNUM_TOT^2, zNUM_TOT^3, zNUM_TOT^4
e. Dependent Variable: SAL_AP

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	414.025	1.770		233.895	.000	410.552	417.498
	zNUM_TOT	36.113	1.788	.518	20.201	.000	32.606	39.621
2	(Constant)	425.741	1.878		226.667	.000	422.056	429.427

	zNUM_TOT	66.782	2.883	.958	23.168	.000	61.126	72.438
	zNUM_TOT^2	-12.402	.951	-.539	-13.036	.000	-14.268	-10.535
3	(Constant)	439.687	2.329		188.791	.000	435.117	444.256
	zNUM_TOT	82.261	3.217	1.180	25.568	.000	75.948	88.574
	zNUM_TOT^2	-39.526	3.000	-1.718	-13.174	.000	-45.413	-33.639
	zNUM_TOT^3	5.096	.537	1.030	9.493	.000	4.043	6.149
4	(Constant)	450.396	2.671		168.609	.000	445.155	455.638
	zNUM_TOT	77.443	3.201	1.111	24.195	.000	71.162	83.723
	zNUM_TOT^2	-73.648	5.350	-3.202	-13.765	.000	-84.146	-63.150
	zNUM_TOT^3	24.375	2.584	4.925	9.431	.000	19.304	29.446
	zNUM_TOT^4	-2.429	.319	-2.447	-7.618	.000	-3.054	-1.803

a. Dependent Variable: SAL_AP

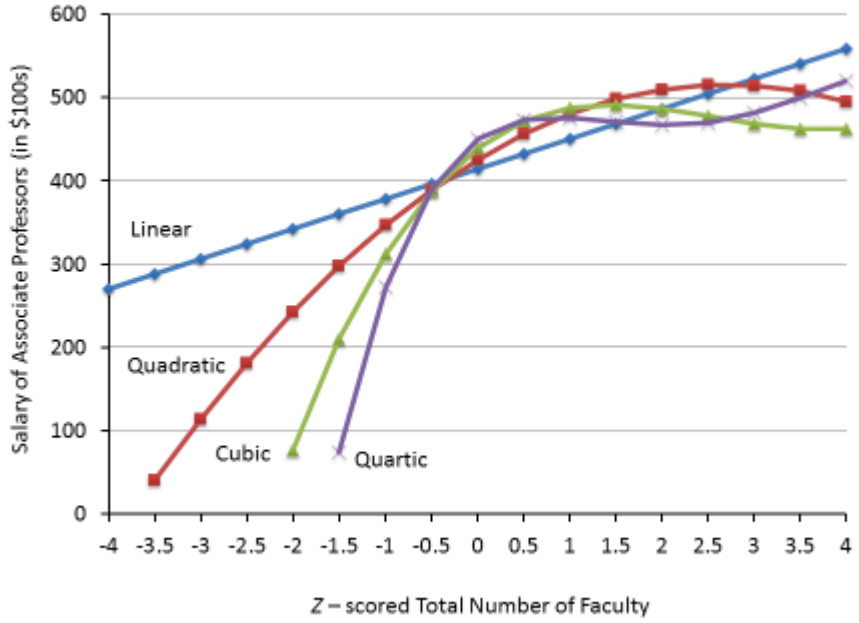

Figure 7.10: Curvilinear effects of institutional size on salary of associate professors (in hundreds of dollars) following data cleaning. Data: AAUP

Additionally, because we are talking about salary as our DV, I deleted negative predicted values from the graph as they did not make sense. As you can see in Figure 7.10, the curves add some nuance to the understanding of the relationship between the two variables. For example, there is a good deal of expected change in the lower end of the range, and then a flattening out at the upper end. Thus, it might be the case that the most gains are in a particular range, which a curvilinear effect can help you identify more clearly. Looking at the final (quartic) curve, the expected average salary is essentially flat between about the mean (0) and +3.0, indicating that most of the change is in the smaller institutions, and once you reach a certain point, the effect of increased institutional size on professor salary is much weaker. This is a very different conclusion from the model with only the linear effect, which would lead us to include that institutional size has a constant effect on salary across the entire range.

Example summary of analysis

If this were my dissertation or study to be published, I would summarize the analyses we have just worked through as follows (eliminating much of the detail you would cover such as sampling and measurement):

> The initial analysis predicting salary from institution size indicated several potential problems. First, there seemed to be substantial heterogeneity, and possible non-linearity. Second, examination of standardized residuals and DfFit indicators of influence suggested eleven cases that were inappropriately influential (DfFit > 33 SD from the mean; ZRE > |3.0|). After they were removed, a curvilinear regression analysis was performed entering the original (linear) effect first, then the squared, cubed, and quartic terms on separate steps.
>
> All four steps led to significant improvement in model variance accounted for, as you can see in Table 7.2. Ultimately, this model is relatively strong, accounting for over 44% of the variance in salary. Furthermore, after data cleaning and entry of curvilinear terms, normality or residuals was improved and assumption of homoscedasticity was met. Thus, all terms were retained, and the regression line equation from the final model was graphed and is presented in Figure 7.10.
>
> As you can see in Figure 7.10, the salary of associate professors is strongly related to institution size (or our proxy for size, the total number of faculty) when the institutions are below average, but for those institutions above the mean, there is little added effect of size on faculty salary.

Reality testing this effect

Some readers might be skeptical that the data really show this dramatic difference in effects below 0 (mean) and above 0. I heartily encourage researchers to perform reality testing of effects to ensure some arithmetic mistake is not responsible for the effect observed. Thus, I performed two very quick follow-up analyses. First, I performed a simple linear regression with our two variables, selecting only cases where zNUM_TOT < 0. Next, I performed another identical analysis selecting only cases where zNUM_TOT > 0. If my graphing and interpretation is correct, the first analysis should show a dramatic and strong positive effect, and the second should show a much weaker and flatter effect. I will not waste space with all the details, but the first regression coefficient was b_1 = 193.77 (beta= 0.56, R^2 = 0.31). The second analysis contained a regression coefficient of b_1 = 11.09 (beta= 0.23, R^2 = 0.05), an impressive contrast that strongly supports the conclusion in the text box above (but not analyses you would necessarily report in a paper or dissertation).

Summary of curvilinear effects in OLS regression

I hope at this point you have been persuaded that it is worthwhile to explore your data for curvilinear effects, that it is not difficult to do this exploration, and that the results can be quite interesting. Using these data, we demonstrated that an initial linear analysis that did not meet assumptions of OLS regression, and that without modeling the nonlinear effects, the conclusions from this model would have been somewhat misleading. However, this situation improved dramatically once curvilinear terms were added to the model. Not only were the assumptions more clearly met, but the effect was more nuanced and interesting. You may not care about curvilinear effects in professor salaries, but if this were another set of variables, like time spent on homework and student achievement (or exercise and well-being), would it not be helpful to understand the range where the most benefit occurs?

Curvilinear logistic regression example: Diabetes and Age

This second example of curvilinear effects is from the National Health Interview Survey of 2010 (NHIS2010), wherein we will predict diagnosis of diabetes from body mass index (BMI) using logistic regression. Most of the same principles from the example of OLS regression will apply here as well, with a some slight modifications. We will keep the discussion of model fit and etc. abbreviated as we just covered logistic regression in the previous chapter.

As you can see from Figure 7.11, BMI is positively skewed. Conversion to *z* – scores would leave the intercept at about 27.69, which is in the "overweight" range according to the Centers for Disease Control.[6] This might not be desirable, so we can center the distribution at 20, a healthier BMI.[7] Centering a variable rather than converting to *z* –score will leave it in its original metric, with the intercept at a more meaningful point. This will be the variable we will also square and cube to create the curvilinear analysis. These terms will be entered into the analysis one at a time like the previous OLS regression analysis.

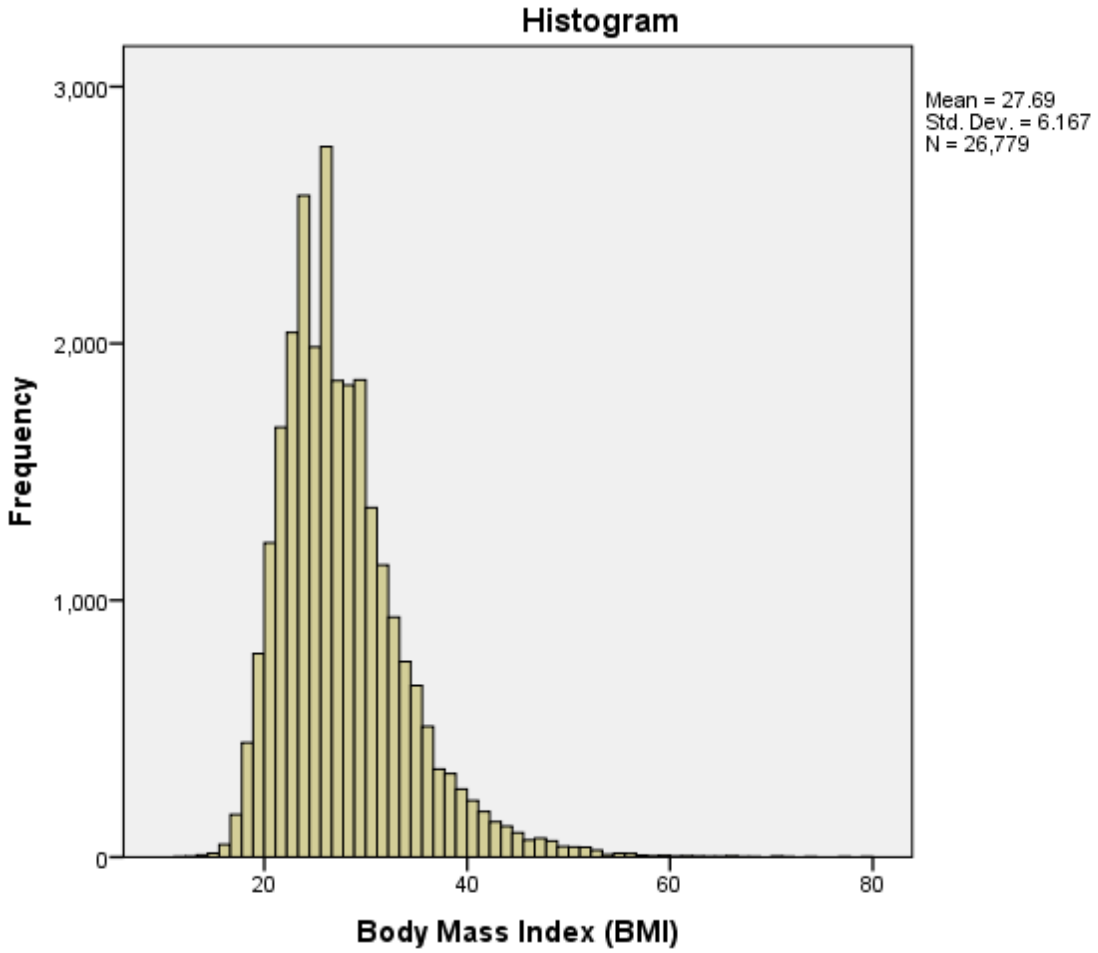

Figure 7.11: Histogram of body Mass Index (NHIS2010).

[6] http://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html
[7] Unfortunately, a BMI I am not likely to see again for a long while…

Remember that in logistic regression, we are looking at improvement of model fit, not variance accounted for. So we are looking for a significant change in -2LL, evaluated as a Chi-Square. Entry of the first term, BMIc (centered BMI) produced a change in -2LL of 986.98 ($p < .0001$). Entry of the quadratic term produced a reduction of -2LL of 114.91 ($p < .0001$), and entry of the cubic term failed to significantly improve the model (Δ-2LL =1.20; $p < .27$). However, this is prior to any data cleaning.

Table 7.3
Relationship of BMI and diabetes

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1	BMIc	.092	.003	1013.633	1	.000	1.096	1.090	1.102
	Constant	-3.071	.037	6872.776	1	.000	.046		
Step 2	BMIc	.169	.008	415.291	1	.000	1.185	1.165	1.204
	BMI2	-.003	.000	98.678	1	.000	.997	.997	.998
	Constant	-3.470	.056	3789.292	1	.000	.031		
Step 3	BMIc	.183	.015	145.901	1	.000	1.201	1.166	1.238
	BMI2	-.004	.001	15.016	1	.000	.996	.995	.998
	BMI3	.000	.000	1.267	1	.260	1.000	1.000	1.000
	Constant	-3.515	.070	2517.835	1	.000	.030		

As you can see from Figure 7.12, there are some cases with some rather extreme standardized residuals (105 of 26779, or 0.39% had a standardized residual greater than 5.0, and were removed). Following removal of these inappropriately influential cases, the model fit became even better than before, as you can see in Table 7.4. You will also see that the cubic term is now significant, although small in effect, and thus will be retained in the model going forward. You might also see that for the final step, I have expanded the number of decimals reported in the regression coefficient column. When dealing with squared and cubed terms and log transformed numbers, increased precision is important. If you examine Table 7.3, you will see 0.000 as the coefficient for BMI^3- which is not really the case and is not really helpful if trying to predict values using that number in a logistic regression equation. Most statistical software will allow you to get more precision either through setting different preferences or by (as in SPSS) clicking on the table itself.

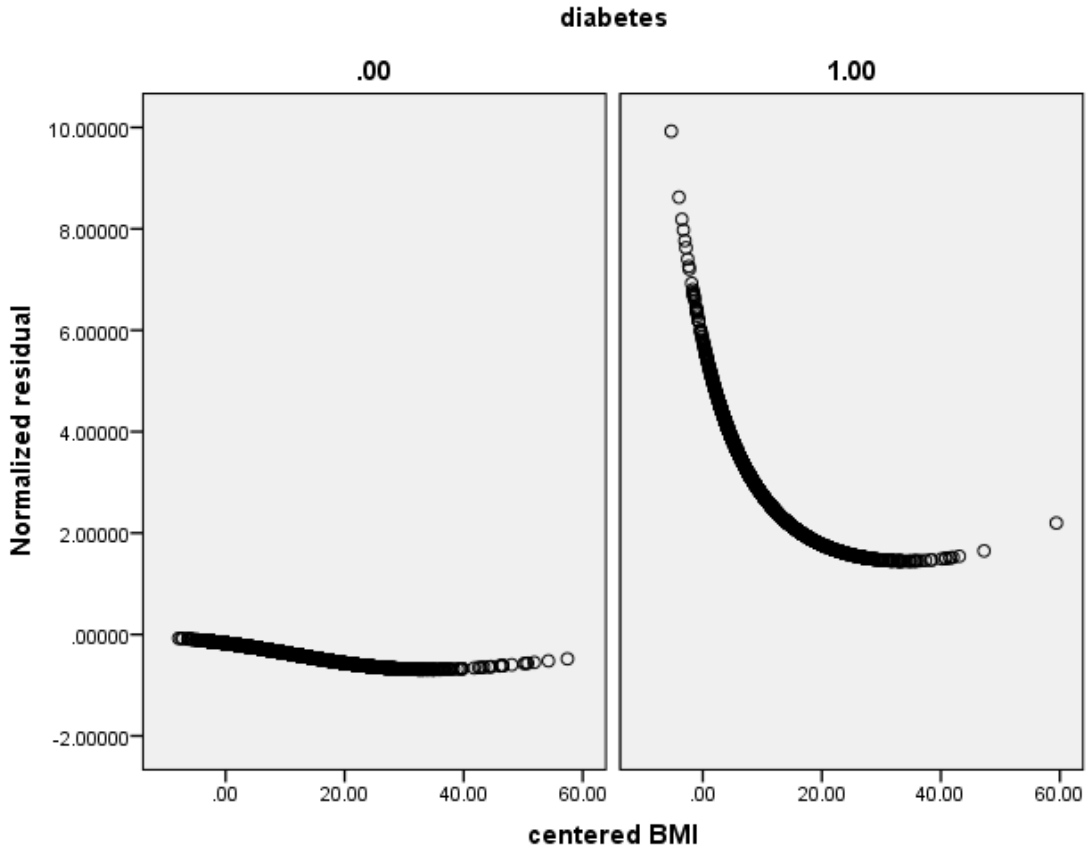

Figure 7.12: standardized residuals from BMI and diabetes analysis.

Table 7.4
Relationship of BMI and diabetes after data cleaning

		Chi-square (Δ -2LL)	df	Sig.
	Step 1	1161.731	1	.000
	Step 2	218.891	1	.000
	Step 3	34.140	1	.000

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1	BMIc	.101	.003	1174.386	1	.000	1.106	1.100	1.112
	Constant	-3.214	.039	6922.436	1	.000	.040		
Step 2	BMIc	.218	.009	540.553	1	.000	1.244	1.221	1.267
	BMI2	-.004	.000	168.583	1	.000	.996	.996	.997
	Constant	-3.840	.064	3614.639	1	.000	.021		
Step 3	BMIc	0.298676	.017	306.258	1	.000	1.348	1.304	1.394
	BMI2	-0.009354	.001	90.563	1	.000	.991	.989	.993
	BMI3	0.000097	.000	37.201	1	.000	1.000	1.000	1.000
	Constant	-4.125529	.084	2416.437	1	.000	.016		

To graph this equation you would create the logistic regression equation from the last step in Table 7.4, as belows:

EQ 7.9:

$$\text{logit}(\hat{Y}) = -4.125529 + 0.298676(\text{BMIc}) - 0.009354(\text{BMI}^2) + 0.000097(\text{BMI}^3)$$

Procedurally, creating predicted logits and conditional probabilities when looking at curvilinear effects is no different than for OLS regression. In this case, recall that we centered BMI at 20, and so 0 will be a BMI of 20, and any increment of ± 1 will be a change in BMI of 1.0 (i.e., we did not convert to z – scores). Furthermore, as discussed in the prior chapter, I encourage readers to convert predicted scores from logit to conditional probability for easier interpretation. I will demonstrate why now.

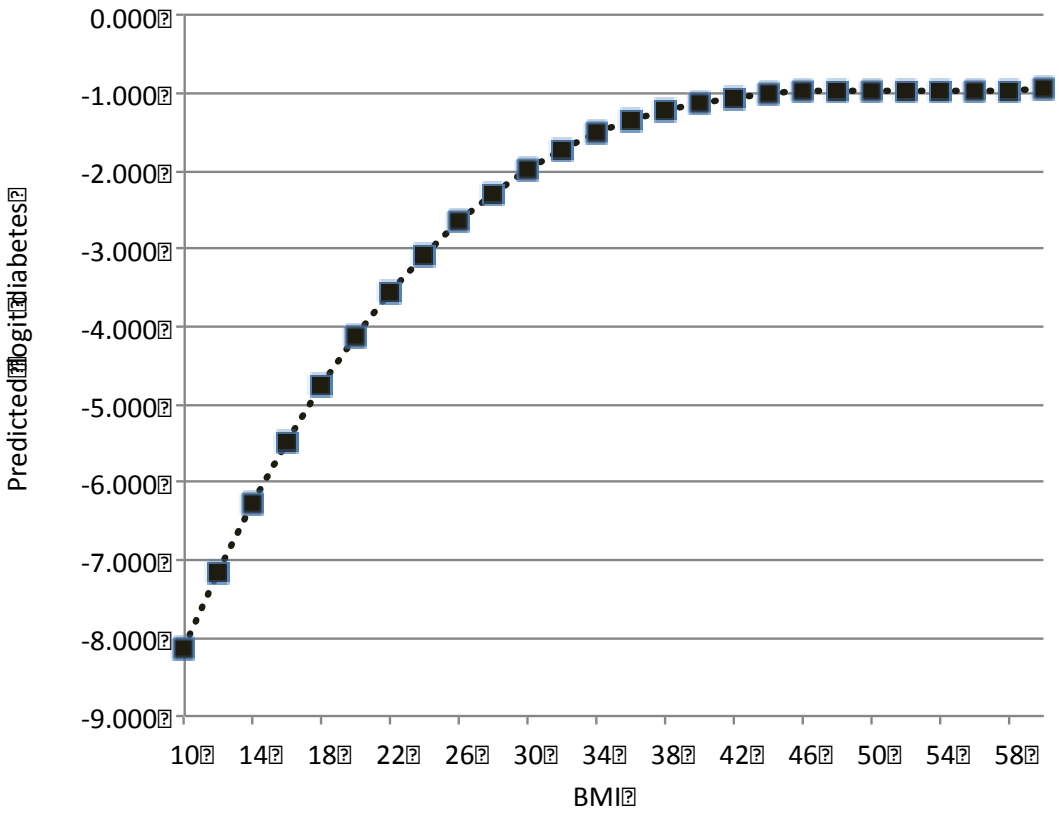

Figure 7.13a: Curvilinear relationship between BMI and diabetes graphed in logits

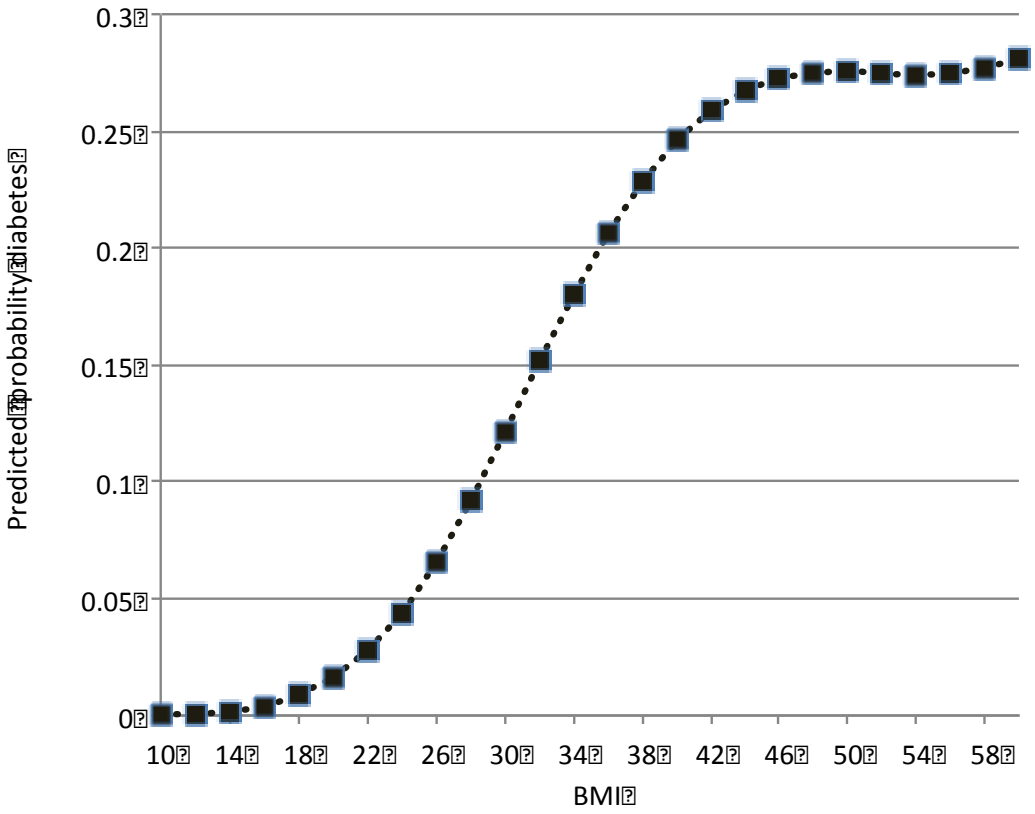

Figure 7.13b: Curvilinear relationship between BMI and diabetes graphed in conditional probabilities

As you can see in Figures 7.13a and 7.13b, the relationship between BMI and diabetes is definitely best defined as a curvilinear effect.[8] However, the nature of logarithms tends to distort the nature of the curve. If you graph the original regression equation, you would conclude that the risk of diabetes is increasing fastest for those with BMI between, say, 10 and 20 (the very skinny and most healthy). However, once these are converted back to conditional probabilities, you can see that this is exactly the wrong interpretation. That is one of the ranges in the relationship where the slope is flattest. As with the example from OLS regression, there are segments of the relationship where there is almost no increase in risk as BMI increases (e.g., between 10-18 and from 46-60) and some areas where each increment in BMI increases the risk of being diagnosed with diabetes dramatically (e.g., 26-42).

[8] also note that despite centering BMI at 20, for the convenience of the reader, I converted BMI back to the original scale when presenting it. Simple touches such as this make the reader's job much easier and reduces the chance of misinterpretation.

An example summary of this analysis

In order to explore the curvilinear relationship between diabetes and BMI, the IV was centered at 20, and then squared and cubed versions of the centered BMI variable were created and entered sequentially (on individual steps) into the analysis. A small (less than 0.4%) of the sample had inappropriate levels of influence by virtue of having standardized residuals greater than |5.0|. After these cases were removed, the entry of each term (linear, quadratic, cubic) contributed to a significant improvement in model fit, as Table 7.4 shows. As all three terms were significant, the final logistic regression equation was used to create predicted values across a broad range of BMI (10-60). These predicted values were converted from logits to conditional probabilities for ease of interpretation.

As you can see in Figure 7.13b, the probability of being diagnosed with diabetes is relatively low and slow to accelerate in adults with low BMI, but begins to rise more rapidly as BMI moves toward the high 20s, and continues increase rapidly until the high 40s, where it levels off at a high prevalence.

Curvilinear effects in multinomial logistic regression

So far we have explored some basic principles around curvilinearity where we have continuous IVs and DVs, and continuous IVs and binary DVs. If you have followed to this point, you should be wondering whether we can apply curvilinear effects to other parts of the Generalized Linear Model: polytomous IVs and DVs. The answer is that because we look for curvilinearity in the IVs, any analysis with a continuous IV is a candidate for potential curvilinearity. This rules out looking for curvilinearity in simple ANOVA type analyses (but not repeated measures!). Let us expand our exploration to multinomial logistic regression, and return to our example from Chapter 6 and the NELS88 data involving student achievement (zACH, our *z* – scored version of the student achievement variable) and marijuana use (MJ; coded 0= never tried it, 1= tried it 1-2 times, 2= tried it 3-19 times, and 3= tried it 20 or more times). In the previous analysis, we observed that higher achievement test scores tended to relate to lower probabilities of trying marijuana at each level.

Let us expand that analysis to add a quadratic term ($zACH^2$) to the equation already described above. The initial model had a likelihood ratio test of $X^2_{(3)} = 129.49$, $p < .0001$ when the first term was entered, with a final -2LL of 12,991.445. Adding zACH and $zACH^2$ significantly improves the model -2LL to 12,969.033, (for a Likelihood Ratio Test of $X^2_{(3)} = 22.41$, $p < .0001$). The cubic effect did not add a significant improvement to model fit, and thus was disregarded.

Table 7.5
Final multinomial model with curvilinear achievement effect predicting marijuana use

Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	13120.935			
Final	12969.033	151.902	6	.000

Parameter Estimates

MJ[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95
								Lo
1	Intercept	-2.104	.040	2813.651	1	.000		
	ZACH	-.221	.032	48.273	1	.000	.801	
	zach2	-.049	.030	2.635	1	.105	.952	

	Intercept	-2.404	.046	2736.645	1	.000	
2	ZACH	-.159	.038	17.068	1	.000	.853
	zach2	-.110	.037	9.030	1	.003	.896
	Intercept	-2.806	.057	2457.751	1	.000	
3	ZACH	-.247	.048	26.506	1	.000	.781
	zach2	-.165	.048	12.033	1	.001	.848

a. The reference category is: 0.

Binary logistic regression models were created to determine if there were any inappropriately influential outliers with the curvilinear effect in the analysis. Examining the deviance residuals, for example, found many cases with values over 2.00, which would be potential candidates for removal. However, none exceeded 3.00, and thus all were retained. As with the logistic regression analysis, the predicted values were restricted to reasonable ranges (in this case, zACH between -2 and +2, which captured most of the sample) and values were converted from logits to conditional probabilities. As you can see in Figure 7.14, the probability of trying marijuana 1-2 times, 3-19 times, or 20+ times (compared to 0 times) remains relatively flat while achievement is below average, and then tends to drop more steeply. The exception is the second group (tried marijuana 1-2 times), which is more linear of an effect, with the probability decreasing more monotonically across the entire range of achievement. This is expected as for this group, the quadratic effect was not significant.

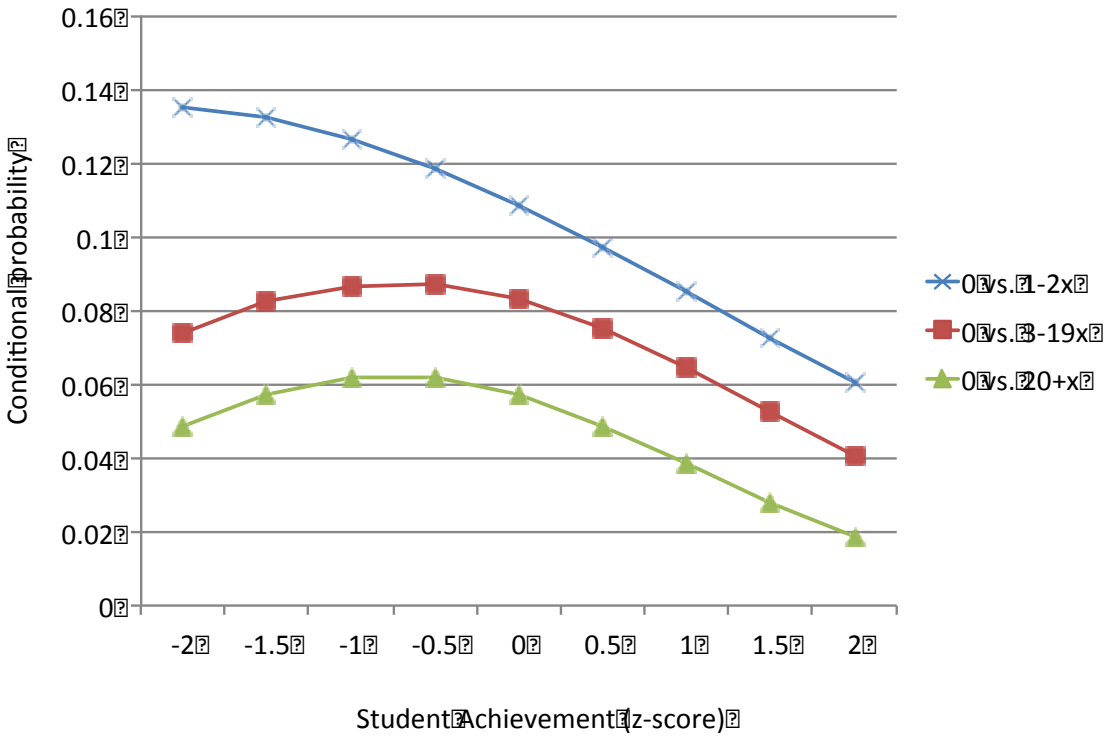

Figure 7.14: curvilinear effect of student achievement on marijuana use. SOURCE:NELS88

Replication becomes important

All effects should be replicated in independent data sets. This is one of the most basic aspects of the endeavor we call science. Without replicability (the ability to produce similar results by performing similar tasks on similar objects) we do not have science. Yet few pursue exploration of replicability. The more complex analyses get, the more important it becomes to replicate our results to ensure that the effects are not merely taking advantage of a peculiar sample. This is particularly important with the smaller samples that are common even in top-tier journals in the behavioral and social (and often, health) sciences. In future chapters we will explore this issue in depth with a variety of effects. For now, let me demonstrate the volatility of curvilinear effects with two random, small samples from the AAUP data we opened the chapter with. Two random samples of approximately 15% (N = 176) will be analyzed in the same way, and the curves compared to see how closely the effect would replicate in two independent samples.

As you can see from Table 7.6, the model summaries show similar patterns, in that all steps are significant, but the details, including the variance accounted for, varies substantially (54% vs. 44%). Additionally, in the second sample, the last two steps are only marginally significant, meaning that had the details been slightly different, the conclusions about the nature of the curve might have been different (quadratic only vs. quartic).

Table 7.6
Replication of two small(er) samples predicting faculty salary from institution size

Model Summary[e]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
Sample 1(N=176)									
1	.530	.281	.277	60.016	.281	67.991	1	174	.000
2	.660	.436	.430	53.295	.155	47.653	1	173	.000
3	.712	.506	.498	50.018	.070	24.408	1	172	.000
4	.732	.536	.525	48.636	.030	10.913	1	171	.001
Sample 2 (N=176)									
1	.553	.306	.302	62.393	.306	73.577	1	167	.000
2	.641	.411	.404	57.662	.105	29.526	1	166	.000
3	.651	.424	.414	57.161	.014	3.925	1	165	.049
4	.663	.440	.427	56.541	.016	4.638	1	164	.033

Sample 1: Final model (Step 4)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
4	(Constant)	462.213	6.504		71.063	.000	449.373	475.052
	zNUM_TOT	89.000	8.495	1.283	10.476	.000	72.231	105.769
	zNUM_TOT^2	-87.862	13.265	-4.136	-6.624	.000	-114.046	-61.678
	zNUM_TOT^3	29.678	7.094	6.536	4.183	.000	15.675	43.682
	zNUM_TOT^4	-3.072	.930	-3.297	-3.303	.001	-4.908	-1.236

Sample 2: Final model (Step 4)

Model	Unstandardized Coefficients	Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B

		B	Std. Error	Beta			Lower Bound	Upper Bound
4	(Constant)	449.811	7.933		56.698	.000	434.146	465.476
	zNUM_TOT	75.028	8.780	1.123	8.545	.000	57.691	92.365
	zNUM_TOT^2	-56.205	15.277	-2.635	-3.679	.000	-86.370	-26.040
	zNUM_TOT^3	19.617	7.840	4.240	2.502	.013	4.136	35.098
	zNUM_TOT^4	-2.217	1.029	-2.303	-2.154	.033	-4.250	-.184

a. Dependent Variable: SAL_AP

As you can see in Figure 7.15, when graphed across a more reasonable range that includes only positive predicted salaries, the curves look similar, although the details differ, particularly in the lower ranges. In this case, the basic conclusions from the analysis were replicated generally. The lesson we will explore in more detail during later chapters is that in many sciences, replication helps establish whether an effect is likely to be found in subsequent samples or whether it was largely an artifact of a particular sample. In this case, a new sample from the same population is likely to give us the same general result, within a certain range.

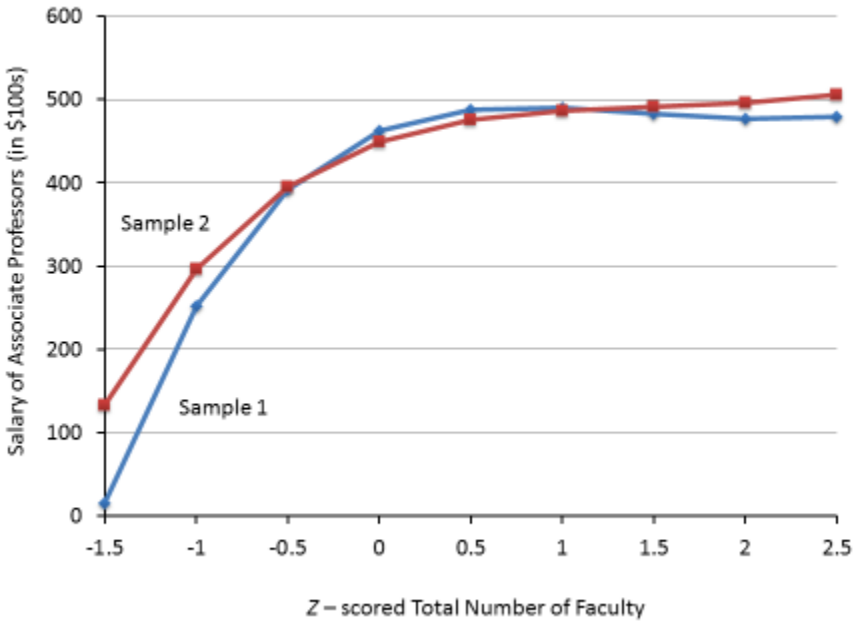

Figure 7.15: replication of AAUP curvilinear analysis in two smaller samples

More fun with curves: Estimating minima and maxima as well as slope at any point on the curve

Although we will explicitly discuss logistic regression in this section because that is the focus of this book, these principles should work with any type of regression. In fact, Aiken and West (1991, see pp. 72-76) explicitly discuss this issue in their excellent treatise on interactions in OLS regression.

Any equation can be manipulated with calculus according to simple rules to allow post-hoc probing of regression line equations. In complex curvilinear equations this can be particularly fun, as you can estimate where the curve reaches a minimum or maximum, or you can estimate the slope at any particular point on the curve to estimate how fast the probabilities are changing.[9]

Those of you who have taken (and remember) basic calculus will remember that taking the first derivative of any equation allows you to estimate slope. So, for example, taking a simple linear equation from the NHIS2010 database, we can look at the logistic regression equation relating AGE and DIABETES (you will be performing these analyses in this chapter's enrichment exercises). The original equation was:

EQ. 7.9a:

$$\text{Logit}(\hat{Y}) = -4.631 + 0.45X$$

or expressed more fully, EQ. 7.9b:

$$\text{Logit}(\hat{Y}) = -4.631X^0 + 0.45X^1$$

Being more specific, the intercept has an x raised to the 0 power, which is 1 (anything raised to the 0 power is 1), and thus it is often eliminated from the regression equation by convention. Further, the X is raised to the first power, and anything raised to the first power is itself. This might seem like more detail than is needed, but once we start adding quadratic and cubic terms, or taking derivatives, this starts to make some sense. For example, the quadratic equation for AGE and DIABETES is

EQ. 7.10a:

$$\text{Logit}(\hat{Y}) = -8.56625X^0 + 0.19402X^1 - 0.001301X^2$$

The simple rules for taking a derivative are that you multiply each term by the exponent of the X, reducing that exponent by 1.[10] The first term will drop out, as anything multiplied by 0 is 0. Thus, taking the derivative of equation 7.9b (or 7.9a), we get Equation 7.9c:

EQ 7.9c:

$$\frac{d(logit(\hat{Y}))}{dX} = (1)0.45X^0$$

which simplifies to:

$$\frac{d(logit(\hat{Y}))}{dX} = 0.45$$

[9] As many authors have pointed out (Aiken & West, 1991 pp. 73-75; DeMaris, 1993), technically what you are estimating is the slope of a line *tangent to* the point where we are estimating the value for the first derivative. For our purposes these concepts are identical.

[10] Unfortunately, we cannot include an entire course in calculus here. Please refer to good calculus references if you are not familiar with this concept.

In other words, because this is a *linear* equation, not a curvilinear equation, the slope is constant across the entire regression: 0.45 logits per increase in X of 1.0. Perhaps not the most surprising or illuminating outcome, but a simple example of a derivative.

Let's move to the curvilinear example. The derivative for the quadratic formula (Equation 7.10a) is (dropping the constant and simplifying):

EQ 7.10b:

$$\frac{d(logit(\hat{Y}))}{dX} = 0.19402 - 2(0.001301X)$$

or

$$\frac{d(logit(\hat{Y}))}{dX} = 0.19402 - 0.002602X$$

Once we have this first derivative, we can look for the point where the slope is 0 (the minimum or maximum) by setting $\frac{d(logit(\hat{Y}))}{dX}$ equal to 0 and solving for X. We get Equation 7.10c:

EQ 7.10c:

$$0 = 0.19402 - 0.002602;$$

by adding 0.002602X to both sides in EQ 7.10c we get:

$$0.002602 = 0.19402;$$

solving for X we get:

$$X = 74.57 \text{ years}$$

Looking at the curve this makes sense, as visually we can see that the curve levels off around that point and then curves downward:

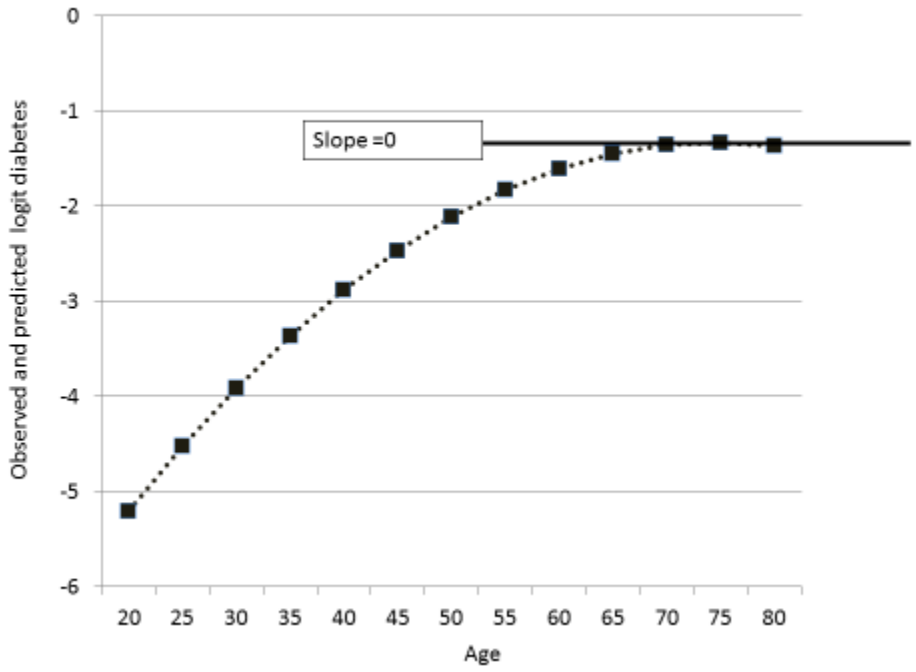

Figure 7.16: Calculating the inflection point of a curve. Source: NHIS2010

Note that we are predicting the slope of logit($\hat{Y}$), and when it reaches zero, that is where the curve has a minimum or maximum and change in direction.[11] However, with the first derivative, we can do more. We can also estimate slopes (in logits) at particular values of X. For example, let us look again at the first derivative of the quadratic equation, and estimate the slope at two other time points (we already know the slope around Age = 75): Age = 25 and Age = 50. By substituting these into the equation, we get slopes of 0.12897 for Age = 25 and 0.06392 for Age = 50. This suggests that the log odds of having diabetes are increasing faster at age 25 than 50. Looking at the graph of logits (Figure 7.16), that seems to hold.

However, looking at the graph of the predicted conditional probabilities (Figure 7.17) it does not. The change in probabilities seems to be much slower at age 25 than age 50. Thus we must be careful to be clear when reporting post hoc probes of these types of analyses, but they can be useful at times. However, note that the extrema is calculated to be identical for both graphs.

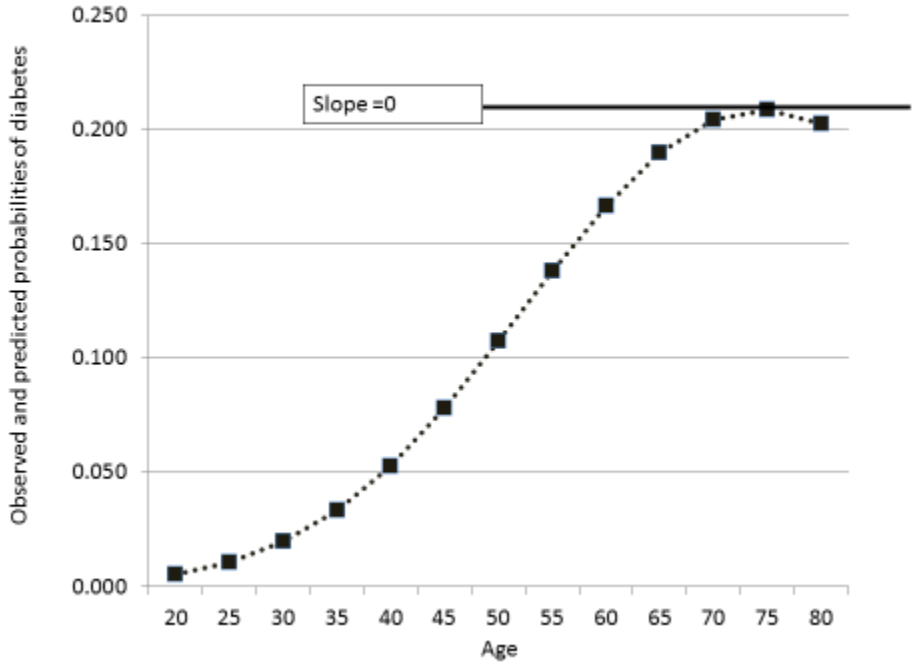

Figure 7.17:The same curve graphed as conditional probabilities. SOURCE: NHIS2010

Let us examine a more interesting curve, that predicting EVERMJ from student achievement. We performed the linear logistic regression analysis for these data in Chapter 5, but if you work through all the enrichment examples at the end of this chapter, you will find there is a cubic curve, meaning it has two points where the slope is equal to 0. The original equation (after data cleaning) is shown below as a spoiler in Equation 7.11a:

EQ. 7.11a:

$$\text{Logit}(\hat{Y}) = -1.1514 + 0.2683(\text{zACH}) - 0.0214(\text{zACH}^2) - 0.3168(\text{zACH}^3)$$

which gives us a first derivative, shown in Equation 7.11b:

EQ. 7.11b:

[11] For you calculus nerds out there, technically we are estimating the slope of a line tangent to the point, but proofs can show that the slope of that line is also the instantaneous slope of our curve at that point.

$$\frac{d(logit(\hat{Y}))}{dx} = 0.2683 - 0.0428(zACH) - 0.9504(zACH^2)$$

If you set Equation 7.11b equal to zero and solve, you find two *extrema*: at -0.55 and at 0.51, both of which seem reasonable given the graph below (graphed in logits rather than predicted probabilities), shown in Figure 7.18:

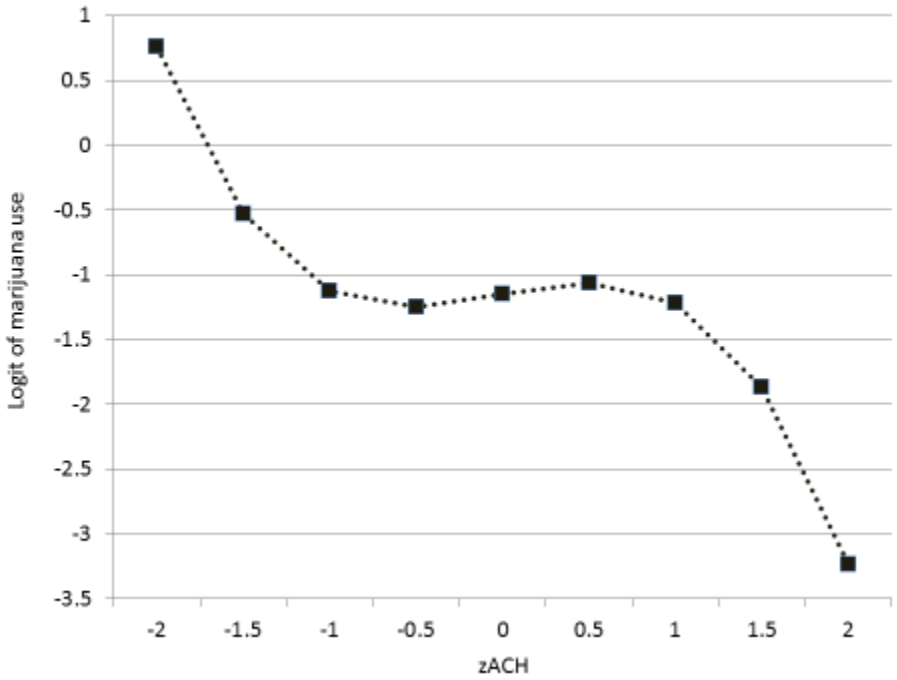

Figure 7.18: Two inflection points for a cubic curve. Source:NELS88

We could again predict slopes at particular points using the first derivative. With this example, let us examine the following three points: -1.75, 0, and 1.75. Substituting into the equation, we get slopes of: -2.57, 0.27, and -2.72, respectively. This tells us that the logits are decreasing relatively steeply in the extremes of the distribution and are relatively flat in the center of the distribution of achievement scores.[12]

In general this procedure should allow reasonable estimation of extrema (minima and maxima) for curves expressed either as logit or probability. Note that while these calculations give us very exact estimates (our diabetes equation has an inflection point at AGE=74.75 years), the precision of estimates through this method is only as good as the data. In curves that replicate well, this might be useful. Where curves are not able to be replicated, this process is really not terribly useful. This is a warning all statisticians using regression or linear modeling need to keep in mind! One can model complex, beautiful curves with poor-quality, biased, or error-filled data and the results are only as good as the ingredients.

[12] There are interesting examples of application of this technique throughout various literatures in science. For example, Boyce and Perrins (1987) used this type of technique of locating extrema to understand and estimate the optimal clutch size for Great Tits (*parus major*, the bird, although I could see how this particular phrasing could lead to confusion) in varying environmental conditions. Apparently there is a curvilinear relationship between clutch size (number of eggs laid) and number of chicks that survive to breed as adults and that this curve is also influenced by whether the year was "bad" or "good" for the birds.

Further, there have been discussions of how to test whether individual point estimates for slope are significantly different from 0. For example, Aiken and West (1991, pp. 77-78) discusses this in regards to OLS regression. I have some reservations about probing the data too much as that (a) increases the risk of over-interpreting the data, unless it is a very large and representative sample, and (b) this too is beyond the scope of this chapter. Perhaps if you encourage all your colleagues and friends to buy the book I will add more of these advanced topics in a second edition!

Summary

This chapter explored how to model curvilinear effects in OLS regression, as well as logistic and multinomial regression. It is relatively simple to find curvilinear effects if you look for them. There are several more examples in the Enrichment section, below.

I had intended to include an example of a curvilinear effect that was due to extreme scores (certainly a possibility!) but was unable to find one in the data sets I was working with. One of the reasons there are so many examples at the end of the chapter relative to other chapters is that as I kept searching for a counter example (removing inappropriately influential scores removed a curvilinear effect) I repeatedly came across relatively powerful and interesting examples of how data cleaning enhanced curvilinear effects. After trying many different modes of data cleaning I failed to find a reasonable example that used appropriate data cleaning to remove a curvilinear effect due solely to inappropriately influential cases. Of course I could manufacture an artificial example, and perhaps I will in the future. At this point, there are two main messages from this chapter.

First, checking analyses for curvilinear effects is not terribly difficult nor is it particularly time-consuming. In a few minutes you can create quadratic and cubic terms for important variables, and in a few seconds an analysis can demonstrate whether there might be a nonlinear effect. Some few minutes more spent data cleaning may amplify or attenuate the effect, and you may end up with a very interesting result.

Second, if you are familiar with simple calculus concepts you can extract interesting details from well-modeled (and replicable) curvilinear equations (such as where the curve flattens out and turns the opposite direction). If you enjoyed this chapter, you will enjoy the chapters to come, in which we have fun with multiple predictors, interactions, and even curvilinear interactions![13]

[13] You may think we were performing analyses that included multiple predictors in this chapter – and in a sense we did, as there were multiple terms being entered as predictors. However, technically, BMI, BMI^2, and BMI^3 are all different aspects of the same variable. So in my mind we were still performing univariate analyses.

Exercises

1. Download the AAUP, NELS88, and NHIS2010 data used in the examples from the chapter and replicate the results from the chapter.
2. Download the EVERMJ data (NELS88) (similar to the data that we explored in Chapter 5) and explore whether there is a curvilinear effect of zACH on EVERMJ.
    a. Perform appropriate data cleaning and tests of assumptions
    b. Summarize results in APA format
    c. Graph effect in conditional probabilities
    d. BONUS: calculate extrema as in the calculus section above to see if your results match mine
3. Within the NHIS2010 data, explore whether AGE predicts DIABETES, whether there are curvilinear effects, etc.
    a. Perform appropriate data cleaning and tests of assumptions
    b. Summarize results in APA format
    c. Graph effect in conditional probabilities
    d. BONUS: calculate extrema as in the calculus section above to see if your results match mine
4. OLS regression example: Within the NHIS2010 data on BMI and Age:
    a. Perform appropriate data cleaning
        i. Both variables have problematic values
    b. Center age at the median (46) and explore whether age predicts BMI.
    c. Graph
        i. Convert centered AGE to actual age to make it easier on the reader but be sure to do the calculations so that you use the centered age.
    d. Summarize in APA format.
    e. BONUS: calculate the age at which average BMI peaks in this population
5. Using the Natality 2013 data (from CDC data on births in the US) on the book web site, predict birth weight (BIRTHWT; measured in grams) from gestational age (GESTWEEK38; number of weeks gestation centered at 38, which is generally considered full-term).
    a. Perform appropriate data cleaning
    b. Graph curve
    c. Summarize in APA format.
6. Using the Natality 2013 data, determine if the amount of weight the mother gains (MOM_WTGAIN) is a significant and important predictor of the birth weight of the infant (BIRTHWT).
    a. Perform appropriate centering etc. of predictor variable, perform appropriate analyses
    b. Graph curve
    c. Summarize in APA format, particularly focusing on effect size (R^2 in particular).

References

Aiken, L. S., & West, S. (1991). *Multiple Regression: Testing and Interpreting Interactions* Thousand Oaks, CA: Sage Publications.

Box, G. E. P., & Tidwell, P. W. (1962). Transformation of the independent variables. *Technometrics*, 531-550.

Boyce, M. S., & Perrins, C. M. (1987). Optimizing Great Tit Clutch Size in a Fluctuating Environment. *Ecology, 68*(1), 142-153. doi: 10.2307/1938814

Cohen, J., Cohen, P., West, S., & Aiken, L. S. (2002). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Mahwah, NJ: Lawrence Erlbaum.

DeMaris, A. (1993). Odds versus Probabilities in Logit Equations: A Reply to Roncek. *Social Forces, 71*(4), 1057-1065. doi: 10.2307/2580130