7

CURVILINEAR EFFECTS IN LOGISTIC REGRESSION

My high school science teacher, Larry Josbeno, was not only a brilliant teacher, but he also was 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[1]:

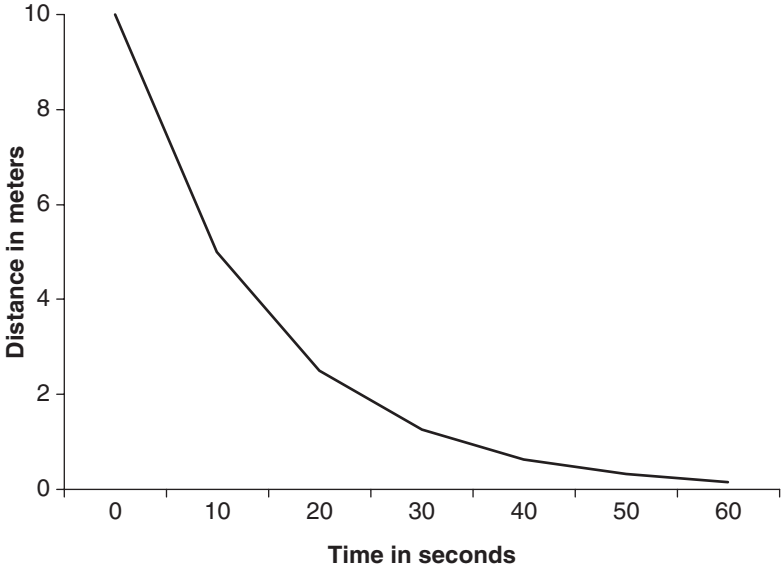
A group of boys are lined up on one wall of a dance hall, and an equal number of girls are lined up on the opposite wall 10 meters apart. Both groups are then instructed to advance toward each other by one half the distance separating them every 10 seconds (i.e., if they are distance d apart at time 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 1 minute they would be close enough for all "practical" purposes.

Enthusiastic adolescent laughter ensued, predictably. Thank you, Mr. Josbeno! But what does this have to do with curvilinear independent variables? 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 shows. And curvilinearity is the topic of this chapter!

In previous chapters, we talked about the assumption that logistic regression is "linear on the logit," meaning that the logits and independent variables are linearly related. I have also asserted that in many areas of science, this assumption may not be tenable. I believe that if we routinely

[1]See Paul Field and Eric W. Weisstein, "Zeno's Paradoxes," from MathWorld—A Wolfram Web Resource (http://mathworld.wolfram.com/ZenosParadoxes.html).

Figure 7.1 Zeno's Paradox in the High School Dance



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 below.

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 logistic regression analyses, and how to graph curvilinear effects.[2]

♦ A BRIEF REVIEW OF THE ASSUMPTION OF LINEARITY

Recall from previous chapters that we assume that the logistic transformation on our binary/categorical dependent variable produces a linear relationship between independent variable(s) and the logit of the dependent variable. If one were to use probit regression (or any other link function

[2]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.

[note we cover probit regression in Chapter 9]), one assumes the relationship will be linear following that transform.

One example in this chapter will be the effect of age on the probability of certain disease states. We will begin by returning to our National Health Interview Survey (NHIS) 2010 (http://www.cdc.gov/nchs/nhis/nhis_2010_data_release.htm) data on diabetes and look at the relationship between age of the patient and the probability of diagnosis with diabetes.

ILLEGITIMATE CAUSES OF CURVILINEARITY ♦

In this chapter, I am most concerned with modeling legitimately curvilinear relationships. As I briefly mentioned in Chapter 4 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).

Violating Equal Intervals in Coding Continuous Variables

It is also possible that poor coding of variables can artificially lead to curvilinearity. Specifically, when researchers take what are conceptually continuous variables (i.e., income, age, or achievement) and create categories for convenience, they might convert ratio or interval measurement to ordinal measurement with unequal categories. I previously introduced the fact that in some government databases, family income is coded unevenly

(such as this example from the Early Childhood Longitudinal Study [ECLS] from the National Center for Educational Statistics [NCES])[3]:

Table 7.1 Total Household Income as Categorized in the ECLS-K Data Set From NCES

1 $5,000 or less	8 $35,001 to $40,000
2 $5,001 to $10,000	9 $40,001 to $50,000
3 $10,001 to $15,000	10 $50,001 to $75,000
4 $15,001 to $20,000	11 $75,001 to $100,000
5 $20,001 to $25,000	12 $100,000 to $200,000
6 $25,001 to $30,000	13 $200,001 or more
7 $30,001 to $35,000	

Data Source: User's Manual for the ECLS-K First-Grade Public-Use Data Files and Electronic Code Book (NCES 2002–135), National Center for Educational Statistics, U.S. Department of Education.

You can see that a continuous variable that could conceptually be ratio measurement (with a true zero point and equal intervals)—family income in dollars—is in this case broken into uneven categories. At the lower end, income is classified into blocks of $5,000 each,[4] while at the top end, the blocks are $25,000, $100,000 each or more. This process can serve to collapse sparse categories or groups into larger groups, convert a linear relationship to curvilinear, or could also be used to convert a curvilinear to linear relationship.[5] To be sure, it seems to violate some of our basic assumptions about measurement and is thus undesirable. One must be careful converting continuous variables to categorical, as we discovered in Chapter 5.

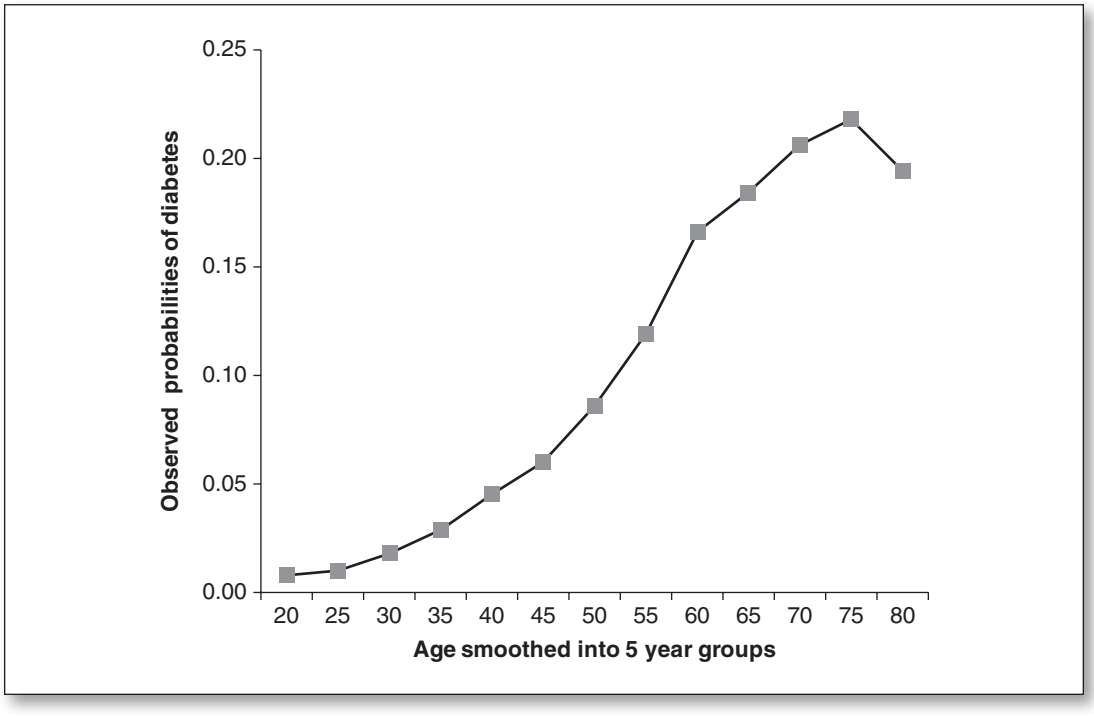
In the case of our first example, age and diabetes, we have a continuous variable (age) that has an interesting curvilinear relationship to the probability of being diagnosed with diabetes, as you can see in Figure 7.2.

[3]http://nces.ed.gov/pubs2002/2002135_2.pdf

[4]In other studies, I have seem more egregious examples of this same thing, where at the lower end income was broken into $1,000 or $2,500 increments, with $100,000 increments at the top end of the scale.

[5]However, as we explored in previous chapters, it is usually better to leave a continuous variable as a continuous variable, transforming if necessary.

Figure 7.2 Age and Probability of Diabetes Diagnosis



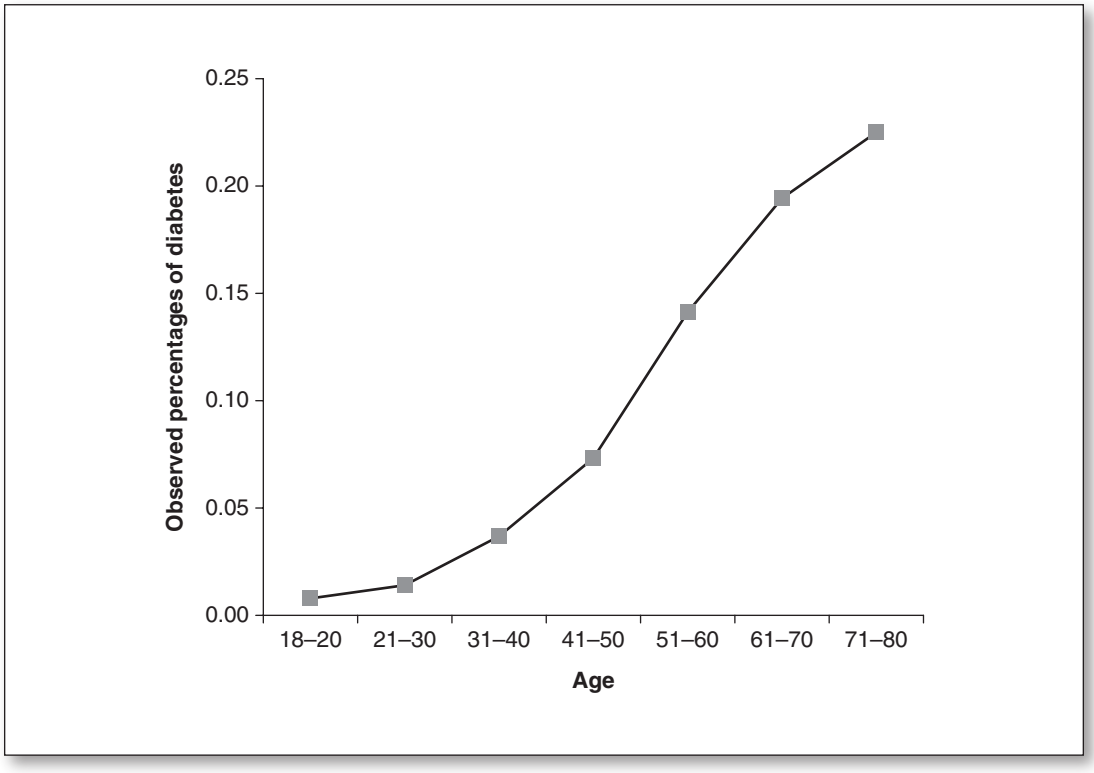
Data Source: National Health Interview Survey of 2010 (NHIS2010), Centers for Disease Control and Prevention.

To follow the previous point, if I group individuals into decades, the curve does change somewhat, although it retains much of general nature, as Figure 7.3 shows. In Figure 7.4, I moved all under 65 into one category and had much smaller age categories above 65. You can see that again changes the nature of the curve.

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.

Figure 7.3 Diabetes and Age Grouped Into Decades

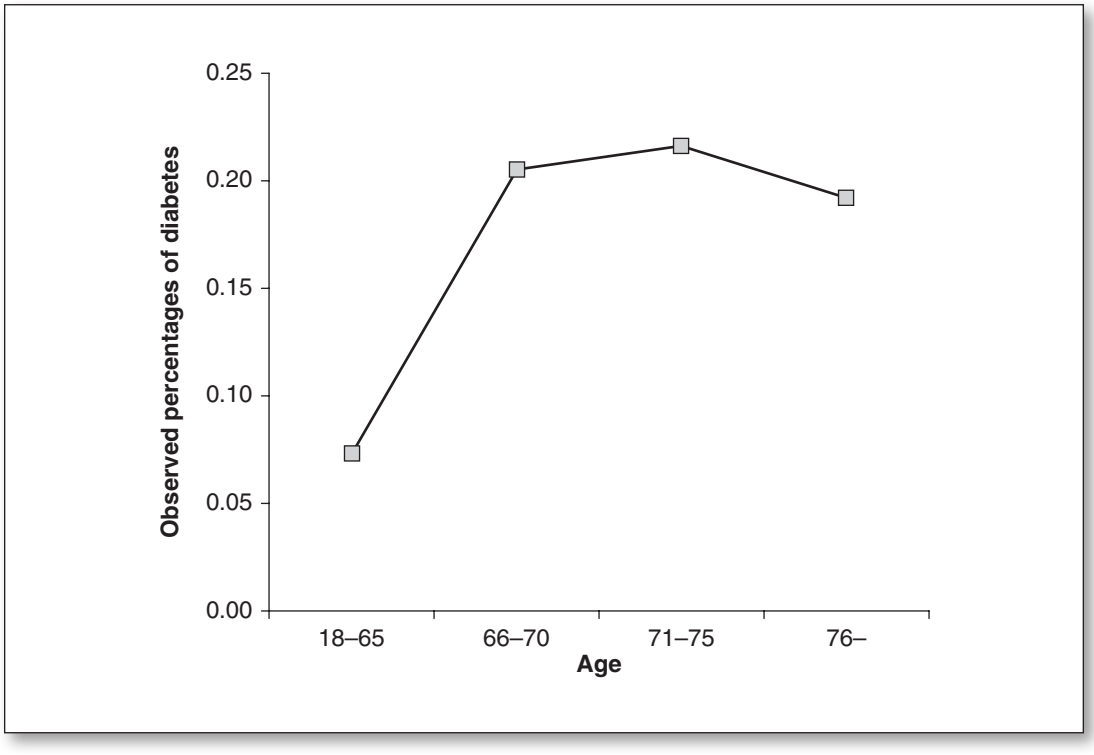


Data Source: NHIS2010, Centers for Disease Control and Prevention.

I suppose it is also possible for highly non-normal data to have the appearance of curvilinearity when in fact it is just another example of poor data cleaning.

Thus, I would argue that prior to examining data for curvilinear effects, one should be sure that the measurement of the variable is defensible (i.e., that you are not creating or masking curvilinearity), that 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.

Figure 7.4 Age and Diabetes in Uneven Categories



Data Source: NHIS2010, Centers for Disease Control and Prevention.

DETECTION OF NONLINEAR EFFECTS ♦

Theory

First and foremost, theory and common sense are always good guides. 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), introduced in Chapter 4 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 nonlinear effects.

The essential process for Box-Tidwell, already described in Chapter 4, 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:

$$V_i = X_i(\ln X_i). \qquad \text{Eq. 7.1.}$$

If the variable V is a significant predictor when X is in the equation, there is a significant curvilinear component to that variable.

$$\text{logit}(\hat{Y}) = b_0 + b_1X_1 + b_2V_1 \qquad \text{Eq. 7.2.}$$

One nice thing about this process is that you then get a good estimate of the nature of the curvilinear effect:

$$\hat{\lambda} = \frac{b_2}{b_1} + 1 \qquad \text{Eq. 7.3.}$$

where b_2 is taken from the second analysis and b_1 is taken from the initial analysis without the V_i in the equation. You can do successive iterations of this process as well, entering $X^{\text{lambda-hat}}$ in place of the original X_i in both the original steps and the calculation of V_i, but in my opinion, that tends to overfit the model unnecessarily. Our data in the social sciences are not the

same character and nature as in the physical sciences and manufacturing, for example.

The final step in this process is to substitute $X^{\text{lamda-hat}}$, where X_i had been in the final analysis. I will also suggest that anytime you incorporate curvilinearity or interactions, you should graph the results for the reader.

CURVILINEAR LOGISTIC REGRESSION ♦ EXAMPLE: DIABETES AND AGE

As a baseline, the unaltered age variable from Figure 7.1 was entered into the logistic regression equation, meaning that the logit and odds ratio reflect increments of 1.0 years, not 1.0 standard deviations as was previously suggested because age in years is a meaningful metric. This model had a −2 log likelihood of 16196.924, which was significant at $\chi^2_{(1)} = 1527.24$, $p < .0001$.

Adding Quadratic and Cubic Terms to the Logistic Regression Analysis

To demonstrate the ad hoc method of exploring curvilinearity, I often add the squared (X^2) and cubed (X^3) terms to the regression equation. With X in the equation, if X^2 is significant that indicates that there is a quadratic (one-bend) curve present in the relationship. With both X and X^2 in the equation, if X^3 is significant, then the curve is cubic (two bends). This procedure, in my experience, captures a reasonable approximation of many curvilinear relationships. Given the generally imprecise nature of

Table 7.2 Relationship of Age and Diabetes—Linear Model Only

		B	*SE*	Wald	*df*	Sig.	Exp(B)	95% CI for Exp(B)	
								Lower	Upper
Step 1[a]	AGE	.045	.001	1381.895	1	.000	1.046	1.044	1.048
	Constant	– 4.631	.074	3890.651	1	.000	.010		

a. Variable(s) entered on step 1: AGE_P.

Data Source: NHIS2010, Centers for Disease Control and Prevention.

measurement in the social sciences (compared with physical or biomedical sciences, or manufacturing, for example), it is important to honor the quality of the data and be careful not to overfit the data beyond what could be expected to generalize.

In this example, both the squared and cubed terms were significant when entered into the equation, but when the cubic term was entered into the equation, it represented a very small ($\chi^2 = 14.14$, $p < .0001$) increment and so was not included in this example for simplicity. The first step was identical to what is reported above. When Age^2 was entered into the equation, the −2 log likelihood was reduced by 298.45 to 15898.47, which was significant at $\chi^2_{(1)} = 298.45$, $p < .0001$.[6]

To graph this equation, you would create the logistic regression equation from Table 7.3:

Table 7.3 Predicting Diabetes From Age and Age^2

		B	*SE*	Wald	*df*	Sig.	Exp(B)	95% CI for Exp(B)	
								Lower	Upper
Step 1[a]	AGE	.19402	.010	405.285	1	.000	1.214	1.191	1.237
	AGE^2	−.001301	.000	253.400	1	.000	.999	.999	.999
	Constant	− 8.56625	.275	971.730	1	.000	.000		

Data Source: NHIS2010, Centers for Disease Control and Prevention.

$$\text{Logit}(\hat{Y}) = -8.56625 + .19402(\text{Age}) - .001301(\text{Age}^2) \qquad \text{Eq. } 7.4^{7}.$$

Procedurally, creating predicted logits and conditional probabilities when looking at curvilinear effects is no different than any simple algebra

[6]Recall that when examining nested models such as this, the difference in the −2 log likelihood is evaluated as a chi-squared statistic with degrees of freedom equal to the number of variables entered on that step. In this case, only one variable was entered, so the chi-squared has one degree of freedom. Most statistical software packages will perform this test for you if you enter the terms on successive steps.

[7]Note that by default SPSS gives only a certain number of decimals—SAS tends to give more. For precision in this sort of application, I like to examine four or five decimals (because we are using logits, in which decimals make a difference!). Double-clicking on the table in the output will allow you to adjust the precision of the output.

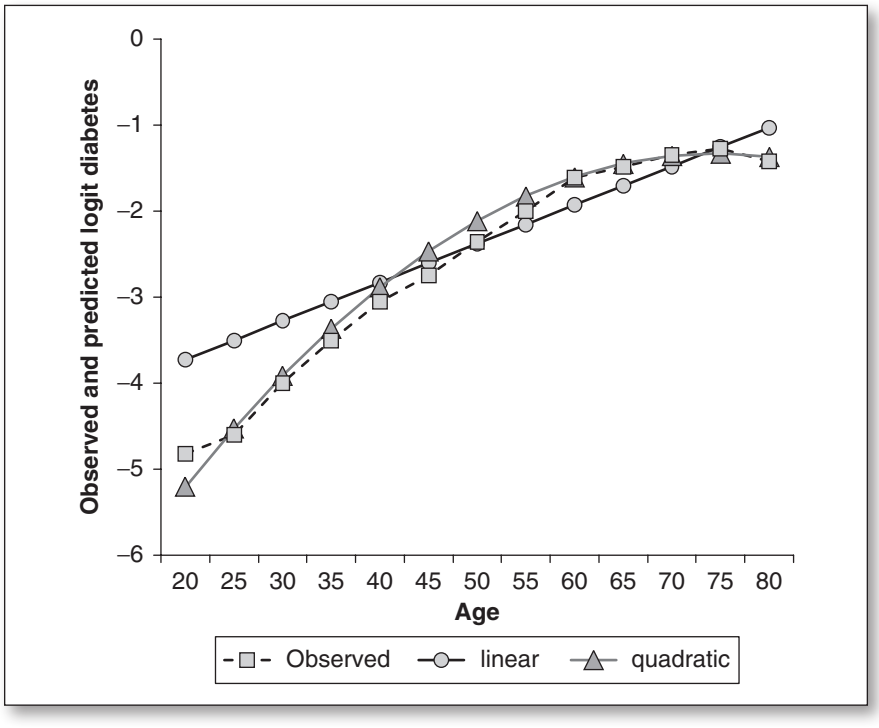
example where you must substitute a value (or several values) of *X* and compute a predicted *Y*. In this case, we can substitute in a range of numbers from 20 to 80, getting predicted logits that can then be converted to predicted probabilities.

As you can see from Figure 7.5, the quadratic line (long dashes with triangles) is a better fit to the actual data (squares with dotted line), although not perfect.

As you can also see from Figure 7.5, graphing the logit (log of the odds of having diabetes in any five-year group) and the predicted logit from the linear and quadratic equations produce similar results in that the curvilinear analysis is much closer to the actual smoothed data than the linear analysis.

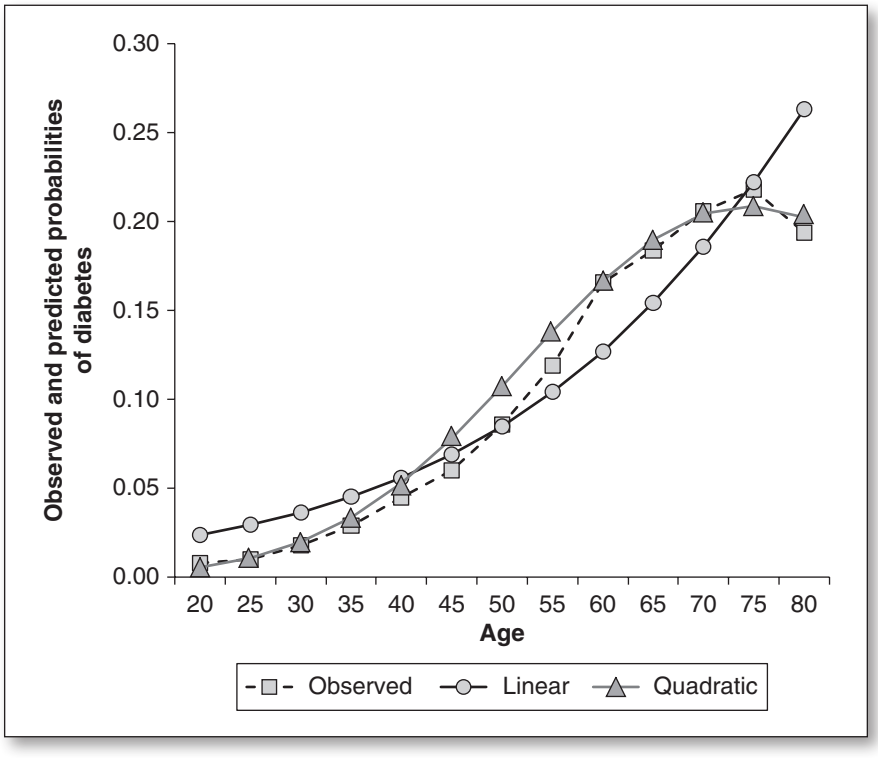
One reason why I like to graph predicted probabilities (as I have advocated for several times in the book already) is that the logit graphs often misrepresent the actual nature of the curve. For example, looking

Figure 7.5 Observed Log Odds of Diabetes, Predicted Logit, and Logit From Quadratic Analysis



Data Source: NHIS2010, Centers for Disease Control and Prevention.

Figure 7.6 Curvilinear Analysis Graphed as Predicted Probabilities Rather Than Logits



Data Source: NHIS2010, Centers for Disease Control and Prevention.

at Figure 7.6, you can see that the observed probability of being diagnosed with diabetes is relatively flat, and then accelerates later in life. Graphing the logit, however, leads to the impression that diabetes rates accelerate early in life and then level out later in life—the opposite of the observed trend.

You can also observe one other interesting thing—remember that logistic regression is "linear on the logit"—meaning that there is assumed to be a linear relationship between the log of the odds of being diagnosed with diabetes and the independent variable (age). In Figure 7.4, when the logit is being graphed, that linear relationship is evident. However, when logits are converted to probabilities, the "linear" relationship is slightly curvilinear. A logarithmic transformation is a nonlinear transform—in this case, taking a nonlinear relationship and making it linear. This highlights the important

reminder that we do *not* assume the relationship is linear between the conditional probabilities of being diagnosed and the independent variable. Only the logit is assumed to be linearly related.

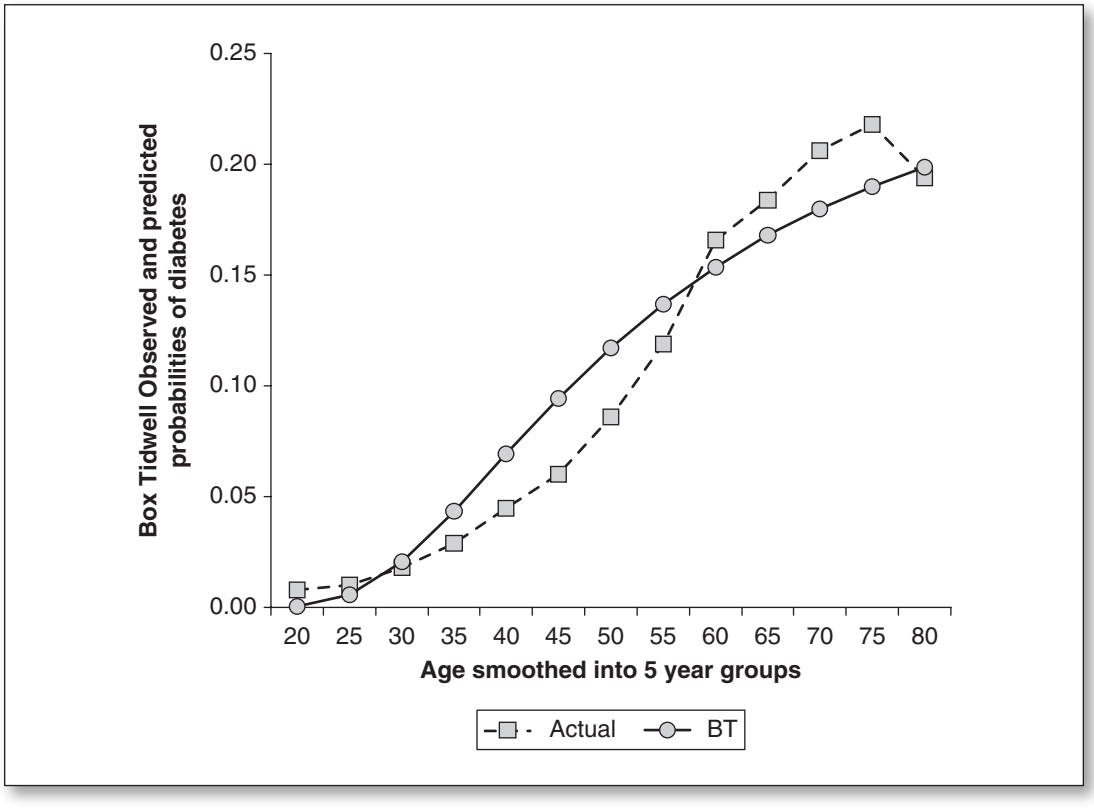
AN EXAMPLE SUMMARY OF THIS ANALYSIS ♦

In order to explore the curvilinear relationship between diabetes and age, squared and cubed versions of the age variable were created and entered sequentially (on individual steps) into the analysis. The linear version of age accounted for significant improvement in the model (−2LL = 16196.92, $\chi^2_{(1)}$ = 1527.24, $p < .0001$). As expected, with only this variable in the analysis, increasing age is associated with increased probability of diabetes ($b = 0.045$, $SE_b = 0.001$, $p < .0001$). When Age^2 was entered into the equation, the −2 log likelihood was reduced by 298.45 to 15898.47, which was significant at $\chi^2_{(1)} = 298.45$, $p < .0001$. (I would summarize both analyses in a single table that combined Tables 7.2 and 7.3 for the convenience of the reader. I would also include a graph similar to Figure 7.6 that simply graphed the curvilinear effect rather than all three lines, which I have mostly included for pedagogical reasons.) As you can see in Figure 7.6, the probability of being diagnosed with diabetes is relatively low and slow to accelerate in relatively young adults, but it begins to rise more rapidly from ages 40–75, at which point it seems to asymptote and then decline slightly.

ESTIMATING CURVILINEAR RELATIONSHIPS ♦ USING BOX-TIDWELL TRANSFORMATIONS

Imagining we were to go about curve estimation more methodically, we could have performed Box-Tidwell initially rather than my old-fashioned ad hoc (and very trustworthy) method. The first step is to estimate the logistic regression equation with the variable of interest in it, as we did earlier. The next step is to create a new version of the variable V that represents $X\ln(X)$ and add it to the equation containing X. The results of this new analysis are interestingly similar to that of the ad hoc analysis earlier when the squared term was entered into the equation as Figure 7.7 illustrates. After entering V, we get a −2 log likelihood of 15914.87, representing an improvement in model fit of $\chi^2_{(1)} = 282.05$, $p < .0001$. According to the

Figure 7.7 Box-Tidwell Transformation of Age



Data Source: NHIS2010, Centers for Disease Control and Prevention.

procedure outlined earlier in the chapter, this significant increment indicates a curvilinear effect, which is not surprising given what we already know of this relationship. We then estimate λ as the ratio of the original regression weight for age divided by the coefficient for *V* plus 1 or:

$$\hat{\lambda} = \frac{b_2}{b_1} + 1 \qquad \text{Eq. 7.5.}$$

I estimate $\hat{\lambda} = (-0.135/.045) + 1 = -2$. When age is transformed in this way and entered into the regression equation, we get −2 log likelihood of 16041.62 (better than the simple linear model but slightly worse than the ad hoc quadratic model above), with a Wald of 819.99, which is slightly better than the combined Wald statistics of the previous analysis.

Table 7.4 Variables in the Equation

		B	*SE*	Wald	*df*	Sig.	Exp(B)	95% CI for Exp(B)	
								Lower	Upper
Step 1[a]	age_BTtransform	−2569.175	89.720	819.991	1	.000	.000	.000	.000
	Constant	−.995	.038	674.019	1	.000	.370		

a. Variable(s) entered on step 1: age_BTtransform.

Data Source: NHIS2010, Centers for Disease Control and Prevention.

Box and Tidwell (and other authors) suggest an iterative approach wherein you substitute the newly transformed variable throughout and perform the regression equation again to see if there is any tweaking that is needed to the lambda. In this case, V and $V\ln V$ were so highly collinear (correlation exceeding 0.99) that it was not possible to perform another iteration.

DATA CLEANING AND CURVILINEAR EFFECTS ♦

Curvilinearity can be caused by unreasonably influential scores, or it can easily be masked by them. Thus, it is important to establish whether a curvilinear effect is being caused by, or masked by, these data quality issues. Turning to our data on marijuana use, we will see an example of how removing 20 out of 540 cases can reveal a curvilinear relationship between marijuana use and student achievement test scores. In this data set (National Education Longitudinal Study of 1988 [NELS88]), all students completed achievement tests, which were combined into single composites at 8th, 10th, and 12th grade. We are using the example of the 8th-grade achievement test score, which has been converted to *z*-scores.

Let us start off with the fact that there is no significant relationship between student achievement test scores at 8th grade and whether the student admitted to using marijuana, as you see in the abbreviated tables (Table 7.5).

Table 7.5 Logistic Regression Equation as Linear, Squared, and Cubic Terms Are Entered Into the Equation

		B	*SE*	Wald	*df*	Sig.	Exp(B)	95% CI for Exp(B)	
								Lower	Upper
Step 1	zBYACH	–.197	.107	3.351	1	.067	.821	.665	1.014
	Constant	– 1.164	.103	126.848	1	.000	.312		
Step 2	zBYACH	–.177	.120	2.167	1	.141	.838	.662	1.060
	zBYACH2	–.039	.102	.149	1	.699	.961	.787	1.174
	Constant	– 1.131	.134	71.266	1	.000	.323		
Step 3	zBYACH	.055	.212	.067	1	.795	1.057	.697	1.602
	zBYACH2	.078	.134	.340	1	.560	1.081	.831	1.407
	zBYACH3	–.132	.101	1.723	1	.189	.876	.719	1.067
	Constant	–1.199	.145	68.669	1	.000	.302		

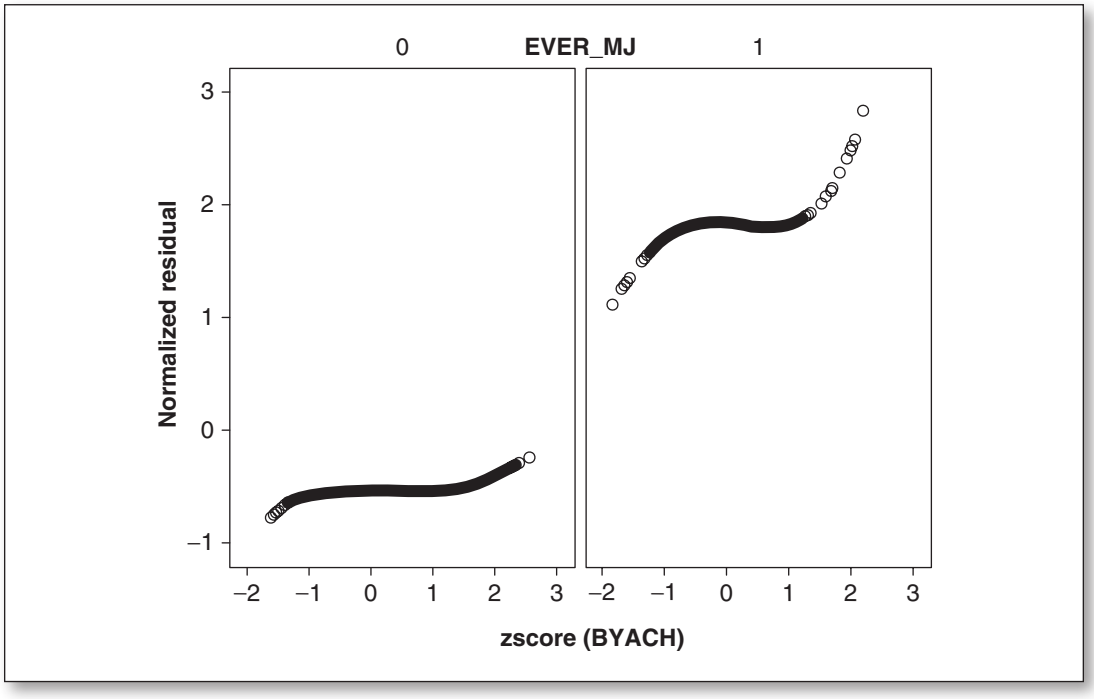
Data Source: NELS88, National Center for Educational Statistics. U.S. Department of Education.

As you can see in Table 7.5, the 95% CIs include 1.0. Although the effect is in the expected direction (students with increasingly high achievement are less likely to admit using marijuana than those with lower achievement test scores), the significance test and CIs do not allow us to conclude there is a significant relationship. Furthermore, adding the squared term for achievement does not improve the situation. In this analysis, neither the linear nor quadratic terms are significant at $p < .05$. Finally, when the cubed term enters the equation, we are left with non-significant results.

After examining some of the typical diagnostic tools available in SPSS (e.g., standardized residuals and DfBetas), the standardized residuals are all within a reasonable range as you can see in Figure 7.8.

However, the DfBetas (shown in Figure 7.9) do seem to have some relatively large values, and so I selected cases with DfBetas for the intercept that were greater than the 1st percentile and less than the 99th percentile. This eliminated 20 of 540 cases, as mentioned earlier but led to a different result. As you can see in Table 7.6, the linear effect is still non-significant (perhaps more so than before).

Figure 7.8 Standardized Residuals From Marijuana and Achievement Analysis



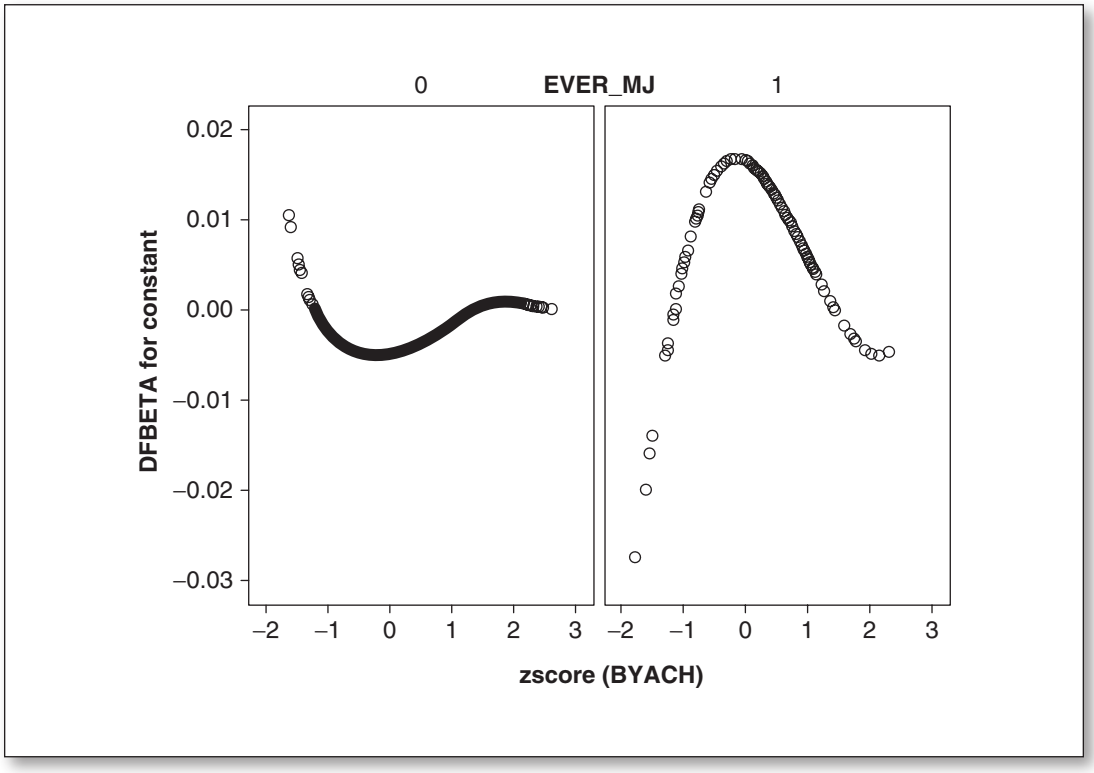
Data Source: NELS88, National Center for Educational Statistics. U.S. Department of Education.

Table 7.6 Variables in the Equation

		B	*SE*	Wald	*df*	Sig.	Exp(B)	95% CI for Exp(B)	
								Lower	Upper
Step 1	zBYACH	–.112	.113	.995	1	.318	.894	.717	1.115
	Constant	–1.350	.111	146.777	1	.000	.259		
Step 2	zBYACH	–.135	.127	1.137	1	.286	.874	.682	1.120
	zBYACH2	.039	.105	.142	1	.706	1.040	.847	1.277
	Constant	–1.384	.146	90.512	1	.000	.250		
Step 3	zBYACH	.296	.227	1.695	1	.193	1.344	.861	2.097
	zBYACH2	.285	.148	3.682	1	.055	1.329	.994	1.778
	zBYACH3	–.247	.110	5.075	1	.024	.781	.630	.968
	Constant	–1.546	.166	87.138	1	.000	.213		

Data Source: NELS88, National Center for Educational Statistics. U.S. Department of Education.

Figure 7.9 DfBetas Graphed by Group



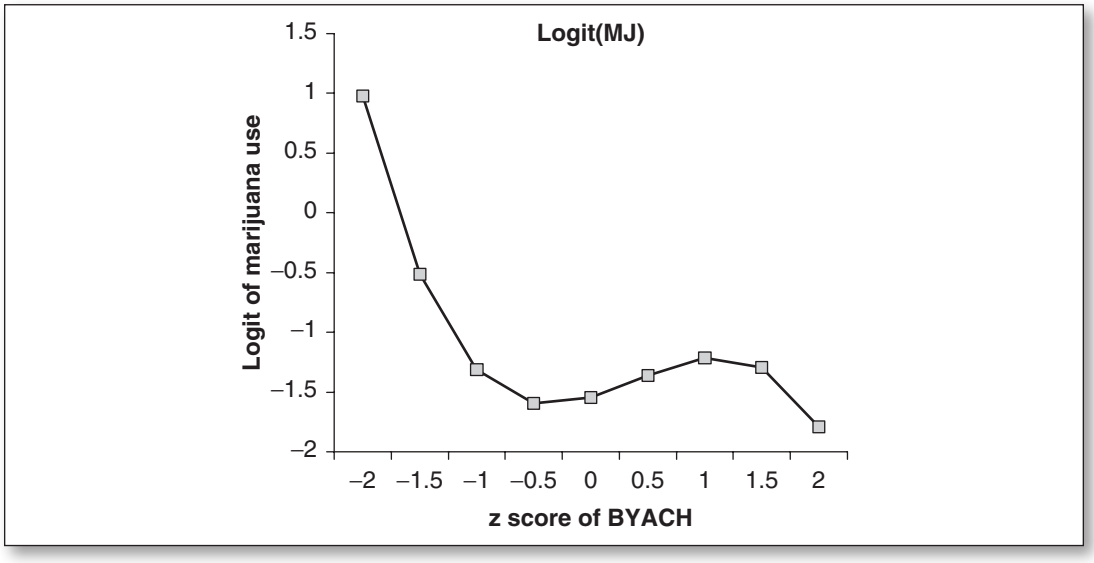
Data Source: NELS88, National Center for Educational Statistics. U.S. Department of Education.

And likewise the quadratic effect is also nonsignificant, but the cubic effect is significant, leaving us with a regression line equation of:

$$\text{Logit}(\hat{Y}) = -1.546 + 0.296(\text{zBYACH}) + 0.285(\text{zBYACH}^2) - 0.247(\text{zBYACH}^3)$$

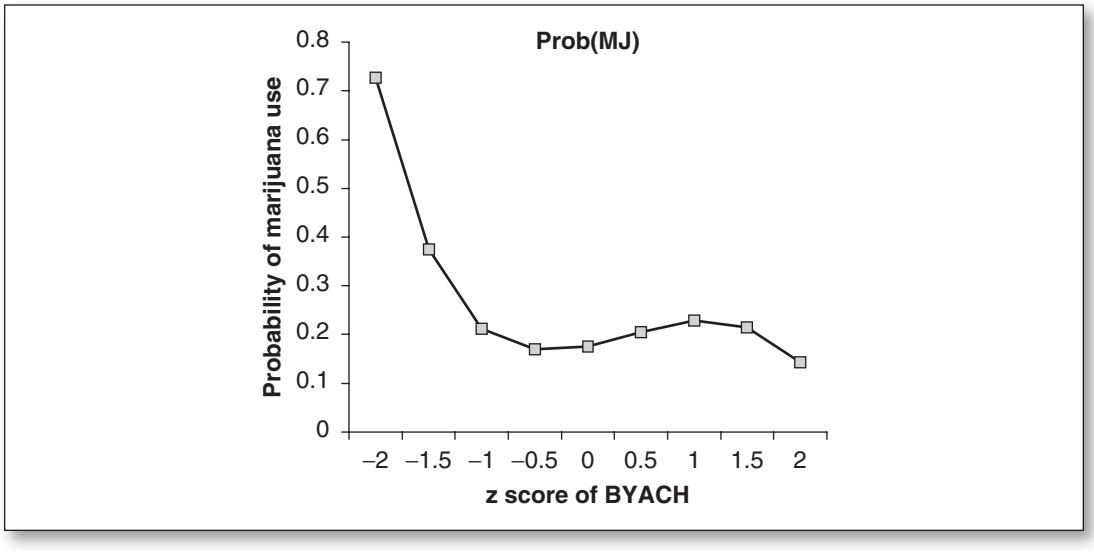
As you can see in Figures 7.10 and 7.11, the curvilinear nature of this relationship is both interesting and somewhat intuitive—and the curve is similar regardless of whether it is expressed in logits or probabilities. In this example, data cleaning revealed an intuitive and interesting curvilinear relationship that was masked by a relatively small number of influential individuals.

Figure 7.10 Curvilinear Relationship Between Marijuana Use and Achievement in Logits



Data Source: NELS88, National Center for Educational Statistics. U.S. Department of Education.

Figure 7.11 Curvilinear Relationship Between Marijuana Use and Achievement in Conditional Probabilities



Data Source: NELS88, National Center for Educational Statistics. U.S. Department of Education.

♦ SAS ANALYSES USING DIFCHISQ

Using the same example and data, we can see that using DIFCHISQ can be equally powerful. Performing the same analysis in SAS, we achieve the same results with noncleaned data:

Table 7.7 SAS Results Prior to Cleaning the Data

Analysis of Maximum Likelihood Estimates					
Parameter	*df*	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	–1.1986	0.1446	68.6692	<.0001
zBYACH	1	0.0551	0.2122	0.0673	0.7953
zBYACH2	1	0.0783	0.1343	0.3399	0.5599
zBYACH3	1	–0.1321	0.1007	1.7229	0.1893

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
zBYACH	1.057	0.697	1.602
zBYACH2	1.081	0.831	1.407
zBYACH3	0.876	0.719	1.067

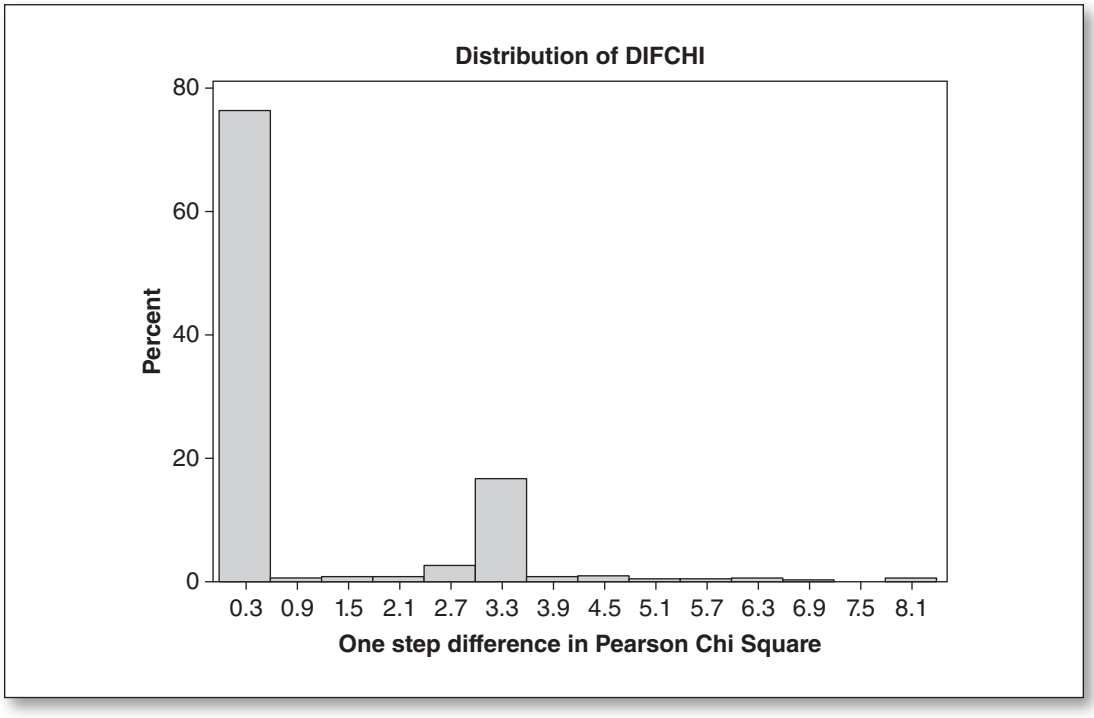
Data Source: NELS88, National Center for Educational Statistics. U.S. Department of Education.

However, looking at the DIFCHISQ results presented in Figure 7.12, we can see that there are a very few cases that seem to have a relatively large influence on the lack of fit for the model.

Removing six cases where DIFCHISQ was greater than 5.0 (remember that about 4 is significant for an χ^2 with one degree of freedom) produced a significant model presented in Table 7.8.

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

Figure 7.12 DIFCHISQ Results From EVER_MJ and zBYACH



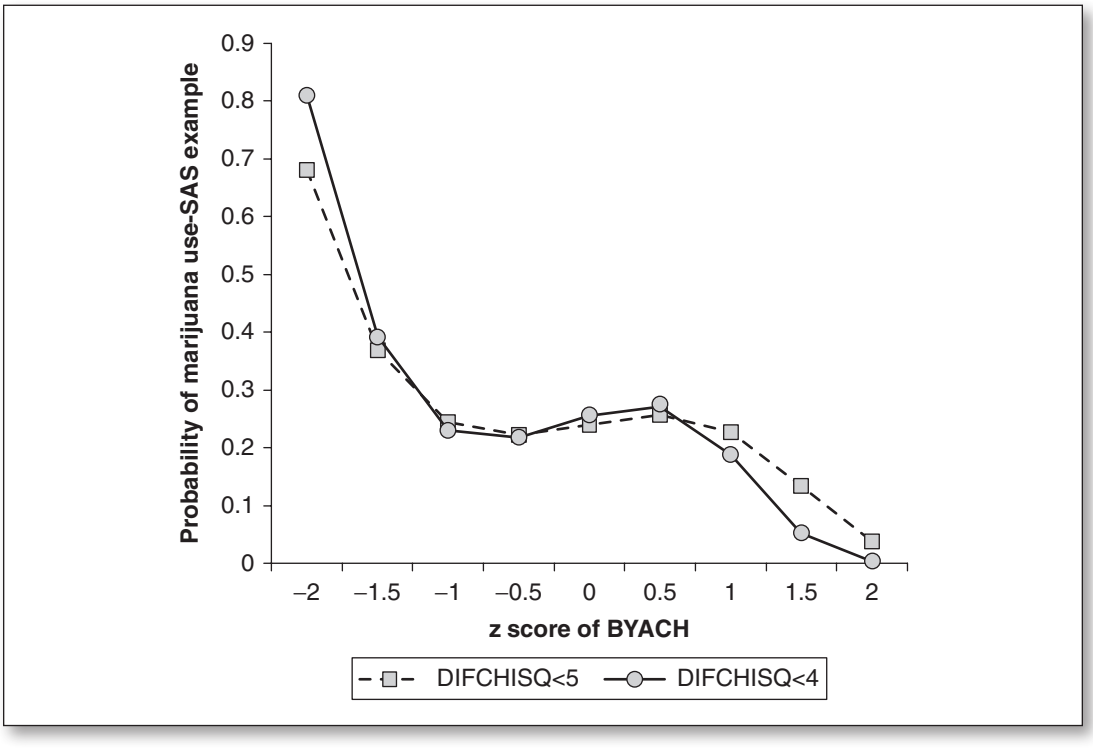
Data Source: NELS88, National Center for Educational Statistics. U.S. Department of Education.

Table 7.8 SAS Results Using DIFCHISQ < 5 to Clean Data

Analysis of Maximum Likelihood Estimates					
Parameter	*df*	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	–1.1514	0.1449	63.1217	< .0001
zBYACH	1	0.2683	0.2421	1.2283	0.2677
zBYACH2	1	–0.0214	0.1401	0.0234	0.8784
zBYACH3	1	–0.3168	0.1314	5.8118	0.0159

Data Source: NELS88, National Center for Educational Statistics. U.S. Department of Education.

Figure 7.13 Curvilinear Relationship Between Student Achievement and Marijuana Use After SAS Cleaning With DIFCHISQ < 5 and < 4



Data Source: NELS88, National Center for Educational Statistics. U.S. Department of Education.

The dashed line in Figure7.13 represents the effect when the data are cleaned by keeping all cases with DIFCHISQ < 5, and the solid line represents a more aggressive cleaning keeping only cases with DIFCHISQ < 4 (removing 6 cases vs. 12 cases out of 540, respectively).

♦ ADVANCED TOPICS IN CURVILINEAR REGRESSION: 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 regression line equation can be manipulated with calculus according to simple rules to allow post-hoc probing. 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.[8]

Those of you who have taken (and remember) basic calculus[9] will remember that taking the first derivative of any equation allows you to estimate slope. So, for example, taking a simple linear equation from our AGE and DIABETES equation, discussed earlier, our original equation was:

$$\text{Logit}(\hat{Y}) = -4.631 + 0.45X \qquad \text{Eq. 7.6.}$$

or expressed more fully:

$$\text{Logit}(\hat{Y}) = -4.631X^0 + 0.45X^1 \qquad \text{Eq. 7.7.}$$

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

$$\text{Logit}(\hat{Y}) = -8.56625X^0 + 0.19402X^1 - 0.001301X^2 \qquad \text{Eq. 7.8.}$$

The simple rules for taking a derivative are that you multiply each term by the exponent of the X, then reduce that exponent by 1. The first term will drop out, as anything multiplied by 0 is 0. Thus, taking the derivative of the first equation, we get:

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

[8]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.

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

which simplifies to:

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

In other words, because this is a *linear* equation, not a curvilinear equation, the slope is constant across the entire regression: 0.45. Perhaps it's not the most surprising or illuminating outcome, but it is a simple example of a derivative. Let's move to the curvilinear example. The derivative for the quadratic formula is (dropping the constant and simplifying):

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

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

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

$0 = 0.19402 - 0.002602X$; by adding $0.002602X$ to both sides we get:

$0.002602X = 0.19402$; solving for X we get:

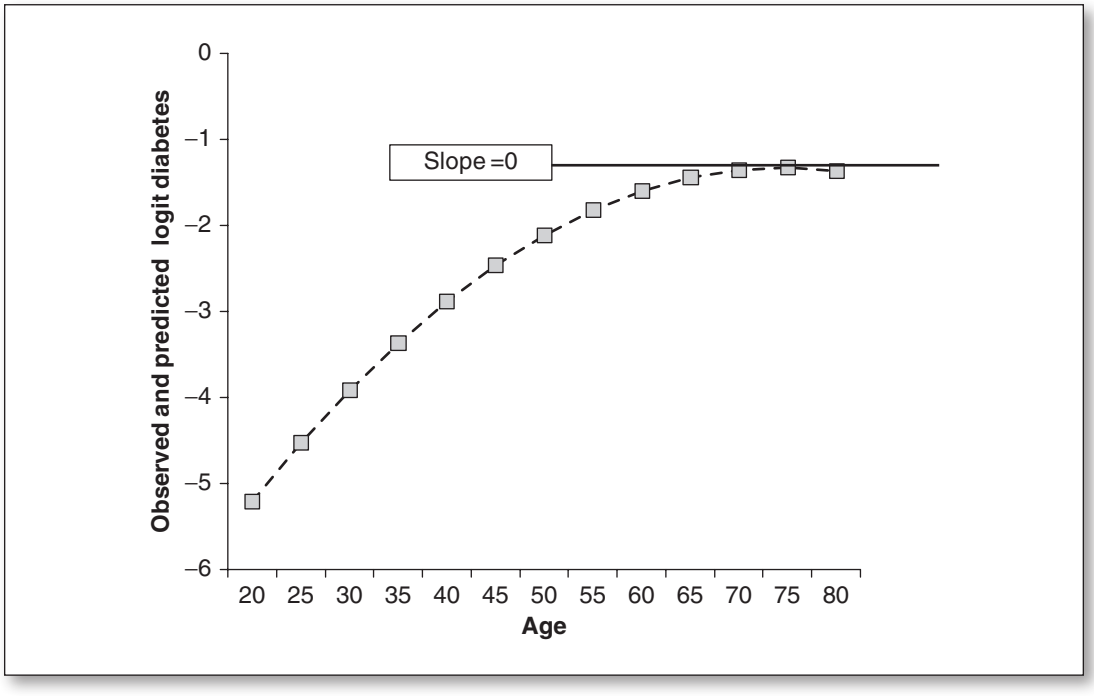
$X = 74.57$ years

Looking at the curve from earlier in the chapter (Figure 7.14) this makes sense, as visually we can see that the curve levels off around that point and then curves downward.

Note that we are predicting the change in the logit($\hat{Y}$). When the logit = 0, that is where the probabilities are 50%, or the odds are 1.0: in other words, there is no difference between the groups, and the slope is 0.

We can also estimate slopes of the tangent lines (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 odds of having diabetes are increasing faster at age 25 than 50. Looking at the graph of logits, that

Figure 7.14 Calculating the Inflection Point of a Curve



Data Source: NELS88, National Center for Educational Statistics. U.S. Department of Education.

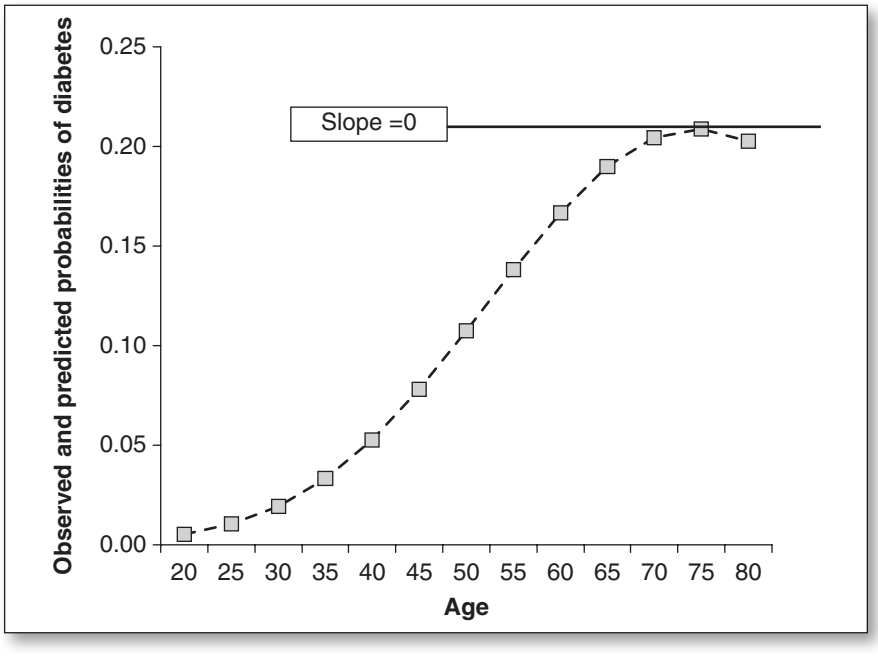
seems to hold. However, looking at the graph of the predicted conditional probabilities (Figure 7.15), it does not. The change in probabilities seems to be much slower at age 25 than age 50. Thus we must be careful to be clear when reporting post hoc probes of these types of analyses, but they can be useful at times.

In the other example, predicting marijuana use from student achievement, the logit and probability curves are similar. This is also a cubic curve, meaning it has two points where the slope is equal to 0. The original equation (after DIFCHISQ = 5 cleaning) was:

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

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

Figure 7.15 The Same Inflection Point Is Apparent When Results Are Graphed as Probabilities



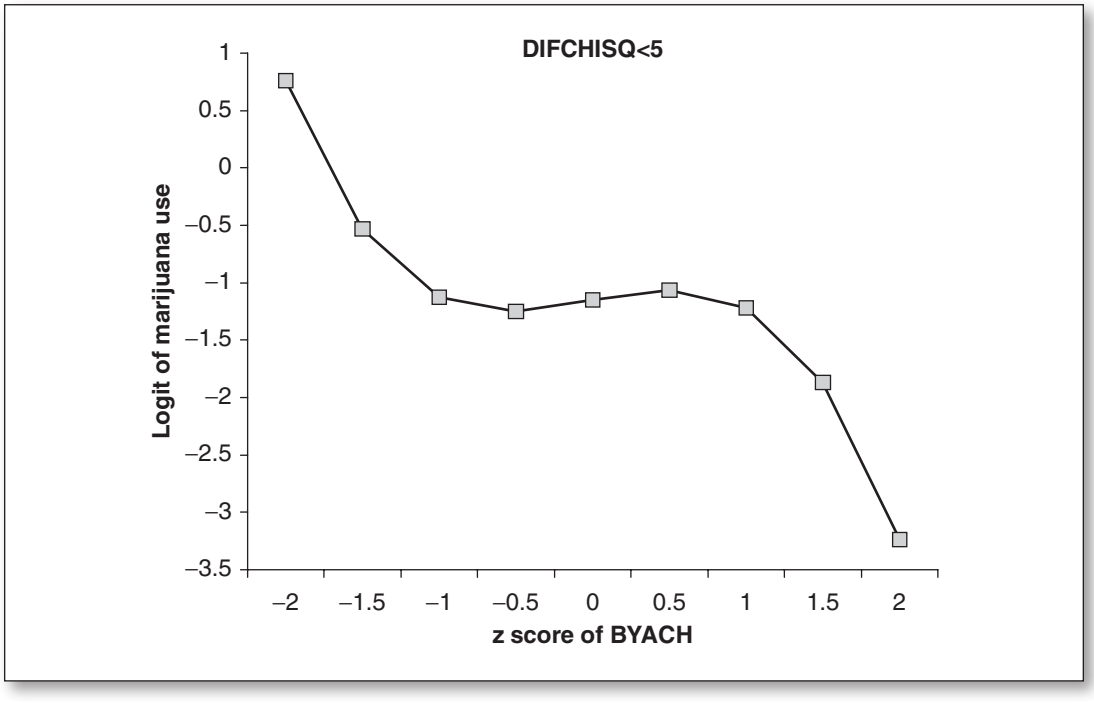
Data Source: NELS88, National Center for Educational Statistics. U.S. Department of Education.

This produces two *extrema:* at −0.55 and at 0.51, both of which seem reasonable given the graph in Figure 7.16 (graphed in logits rather than predicted probabilities).

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.[10]

[10]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 this curve is also influenced by whether the year was "bad" or "good" for the birds.

Figure 7.16 Calculating Inflection Points in a Cubic Curve



Data Source: NELS88, National Center for Educational Statistics. U.S. Department of Education.

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. 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.

The question of expressing slopes in terms of probabilities and more advanced statistical and calculus prospects are beyond the scope of this chapter.

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) discuss this in regards to OLS regression. I have some reservations about probing the data too much, as that (a) increases the risk of overinterpreting 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 logistic regression. As I have shown, it is relatively simple to find curvilinear effects. There are several more examples in the Enrichment section. This chapter became more focused on data cleaning than I had originally planned due to the number of examples I came across that highlighted the efficacy of simple, very conservative data cleaning in revealing or strengthening curvilinear effects. In fact, 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 (standardized residuals, different DfBetas, DIFCHISQ, etc.), I failed to find a reasonable example that used appropriate data cleaning to remove a curvilinear effect. 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, the unintentional message of this chapter reinforces the message from Chapter 4—that cleaning your data in very simple, conservative ways can lead to surprising and unexpected results. I encourage you to explore the data you have, particularly if you have not looked for curvilinear effects before.

If you are familiar with simple calculus concepts, you can glean interesting details from well-modeled curvilinear equations (such as where the curve flattens out and turns the opposite direction). If you enjoyed this

chapter, you will enjoy the next chapter, where we get into multiple predictors, interactions, and even curvilinear interactions![11]

ENRICHMENT

1. Download the data and reproduce the age and diabetes analyses in the chapter. Perform routine data cleaning and compare your results following data cleaning with those reported in the chapter.
2. Using the same diabetes data as above, perform a curvilinear analysis on body mass index (BMI). After performing this analysis, clean the data and perform the analysis on the cleaned data.
3. Using the ELS2002 data predicting dropout/retention from 10th-grade math achievement test scores, perform an analysis examining any curvilinear effects of math achievement on retention. Clean the data, and perform the analysis on cleaned data. Does the result hold?
4. Using the ELS2002 data, perform a curvilinear logistic regression analysis of dropout/retention from family socioeconomic status (zBYSES). Does the curvilinear effect hold after data cleaning?
5. The NELS88 data regarding marijuana use and achievement scores are on the web site for the book.

ANSWER KEY

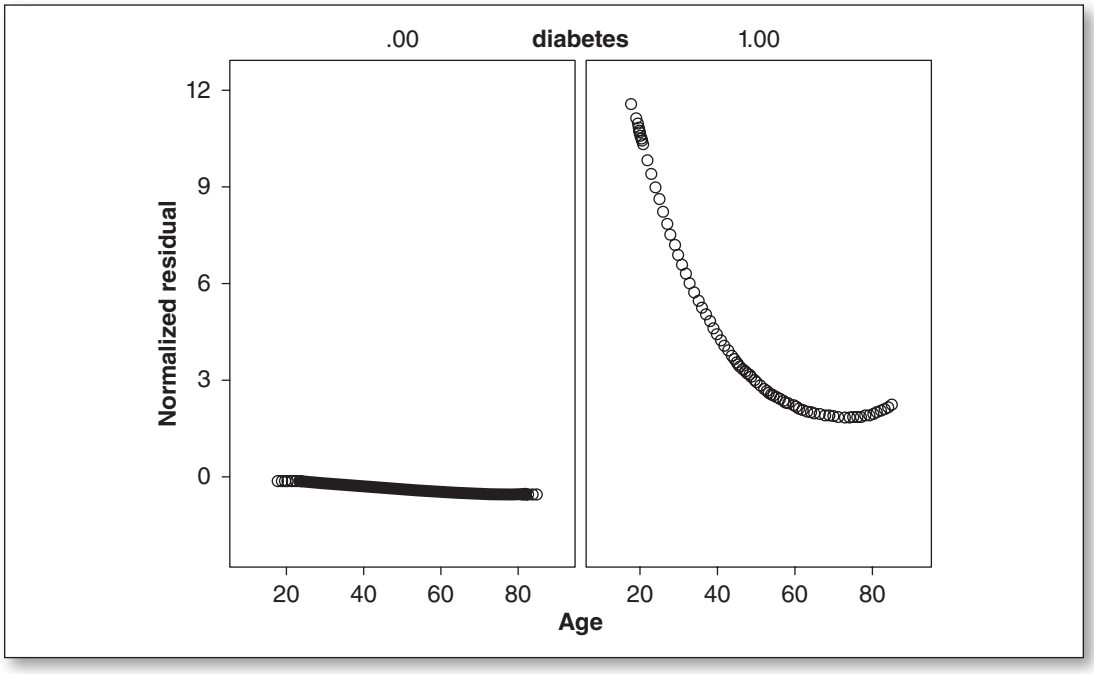
1. Results of Age and Diabetes Analyses Following Routine Data Cleaning

As with the next example (BMI and DIABETES), the model becomes stronger with a bit of data cleaning. For example, examining the standardized residuals from this analysis, we see standardized residuals up to 11.

Eliminating 178 cases with standardized residuals over 5 (to be very conservative) produces improved model fit. As you can see in the abbreviated results, below, the −2LL is reduced 1839.84 with the uncleaned data

[11]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.

Enrichment Figure 7.1 Standardized Residuals From AGE and DIABETES Analysis



Data Source: NHIS2010, Centers for Disease Control and Prevention.

and 2581.065 once a small fraction of the cases with extreme residuals are removed. While these numbers are not from nested models and thus not directly comparable, they are instructive.

Enrichment Table 7.1 Original Model Without Data Cleaning

Omnibus Tests of Model Coefficients				
		Chi-square	*df*	Sig.
Step 1	Step	14.140	1	.000
	Block	14.140	1	.000
	Model	1839.838	3	.000

Data Source: NHIS2010, Centers for Disease Control and Prevention.

Variables in the Equation									
		B	*SE*	Wald	*df*	Sig.	Exp(B)	95% CI for Exp(B)	
								Lower	Upper
Step 1[a]	age	.041	.040	1.040	1	.308	1.042	.963	1.128
	age2	.002	.001	4.405	1	.036	1.002	1.000	1.003
	age3	.000	.000	14.577	1	.000	1.000	1.000	1.000
	Constant	−6.063	.689	77.384	1	.000	.002		

a. Variable(s) entered on step 1: age3

Data Source: NHIS2010, Centers for Disease Control and Prevention.

Enrichment Table 7.2 Cleaned Data

Omnibus Tests of Model Coefficients				
		Chi-square	*df*	Sig.
Step 1	Step	27.416	1	.000
	Block	27.416	1	.000
	Model	2581.065	3	.000

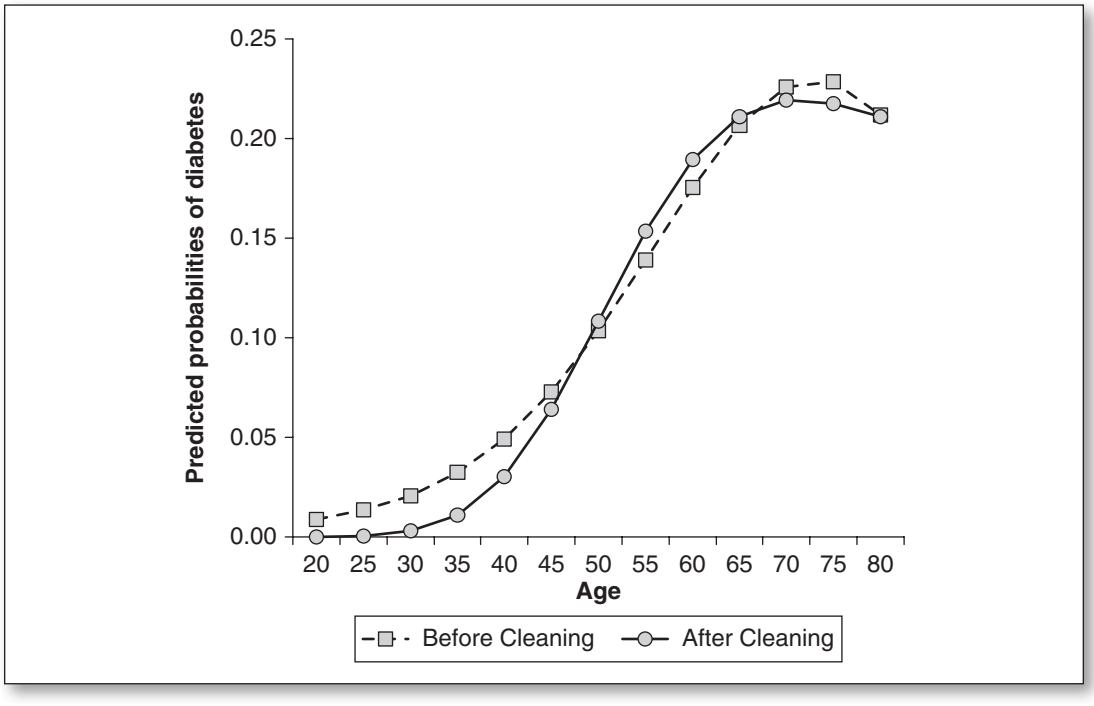
Data Source: NHIS2010, Centers for Disease Control and Prevention.

Variables in the Equation									
		B	*SE*	Wald	*df*	Sig.	Exp(B)	95% CI for Exp(B)	
								Lower	Upper
Step 1[a]	age	.740	.085	76.432	1	.000	2.096	1.776	2.475
	age2	−.009	.001	43.089	1	.000	.991	.988	.993
	age3	.000	.000	24.682	1	.000	1.000	1.000	1.000
	Constant	−20.590	1.634	158.700	1	.000	.000		

a. Variable(s) entered on step 1: age3.

Data Source: NHIS2010, Centers for Disease Control and Prevention.

Enrichment Figure 7.2 Comparison of AGE and DIABETES Before and After Data Cleaning



Data Source: NHIS2010, Centers for Disease Control and Prevention.

As you can see in Enrichment Figure 7.2, I modeled the cubic terms for the original data despite it being a very small effect to maintain parallel analyses. This figure also highlights that a bit of data cleaning in these data produces a more pronounced effect of age—which matches what the model statistics indicate.

Despite the fact that I focused on standardized residuals of a certain extreme magnitude, other indicators (DfBetas, DIFCHISQ) should produce similar results.

2. Results for BMI and DIABETES

The original analyses with all cases is promising: there is a significant overall effect, a significant linear effect, and a significant quadratic (but not cubic effect).

Recall the simple linear analysis for these two variables, shown in Enrichment Table 7.3.

Enrichment Table 7.3 Initial Results for BMI and Diabetes

Omnibus Tests of Model Coefficients				
		Chi-square	*df*	Sig.
Step 1	Step	986.981	1	.000
	Block	986.981	1	.000
	Model	986.981	1	.000

Data Source: NHIS2010, Centers for Disease Control and Prevention.

Variables in the Equation									
		B	*SE*	Wald	*df*	Sig.	Exp(B)	95% CI for Exp(B)	
								Lower	Upper
Step 1[a]	BMI	.092	.003	1013.633	1	.000	1.096	1.090	1.102
	Constant	−4.901	.090	2939.412	1	.000	.007		

a. Variable(s) entered on step 1: BMI.

Data Source: NHIS2010, Centers for Disease Control and Prevention.

This indicates that there is a strong linear effect. However, if we suspect that there is a curvilinear effect, we can enter the squared a cubed terms to explore whether that is tenable. Without cleaning the data, the results are interesting—entering the squared term results in a significant improvement in model fit over simply the linear analysis:

Enrichment Table 7.4 BMI and Diabetes Curvilinear Effect

Omnibus Tests of Model Coefficients				
		Chi-square	*df*	Sig.
Step 1	Step	114.907	1	.000
	Block	114.907	1	.000
	Model	1101.888	2	.000

Data Source: NHIS2010, Centers for Disease Control and Prevention.

(Continued)

Enrichment Table 7.4 (Continued)

Variables in the Equation									
		B	*SE*	Wald	*df*	Sig.	Exp(B)	95% CI for Exp(B)	
								Lower	Upper
Step 1[a]	BMI	.273947	.019	218.858	1	.000	1.315	1.268	1.364
	BMI2	–0.002613	.000	98.678	1	.000	.997	.997	.998
	Constant	–7.904	.316	627.524	1	.000	.000		

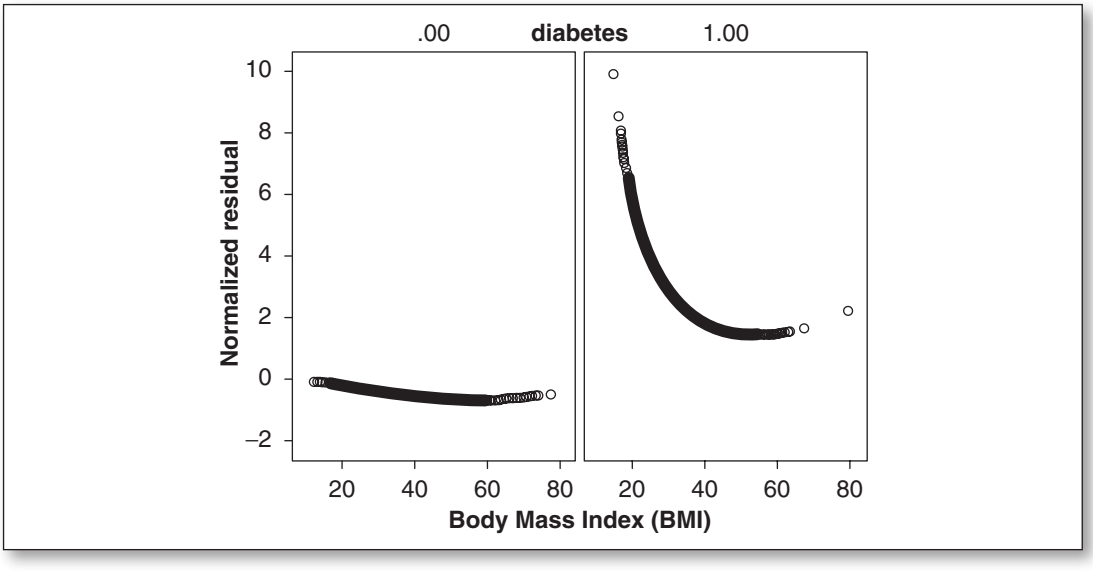
a. Variable(s) entered on step 1: BMI2. Note that I have asked SPSS to provide more decimals than routine to get better precision. You can double-click on the output and get the same precision of results, which is important when graphing these types of outcomes.

Data Source: NHIS2010, Centers for Disease Control and Prevention.

In this analysis, the cubic term does not significantly improve the model ($\chi^2 = 1.197$, $p < .27$).

Examining standardized residuals from this analysis revealed some cases with standardized residuals over 9.0, clearly extreme scores, as Enrichment Figure 7.3 shows:

Enrichment Figure 7.3



Data Source: NHIS2010, Centers for Disease Control and Prevention.

Removing cases with standardized residuals over 4.0 takes our sample from 26,779 to 26,407, removing 372 cases, just over 1%. However, the results are striking:

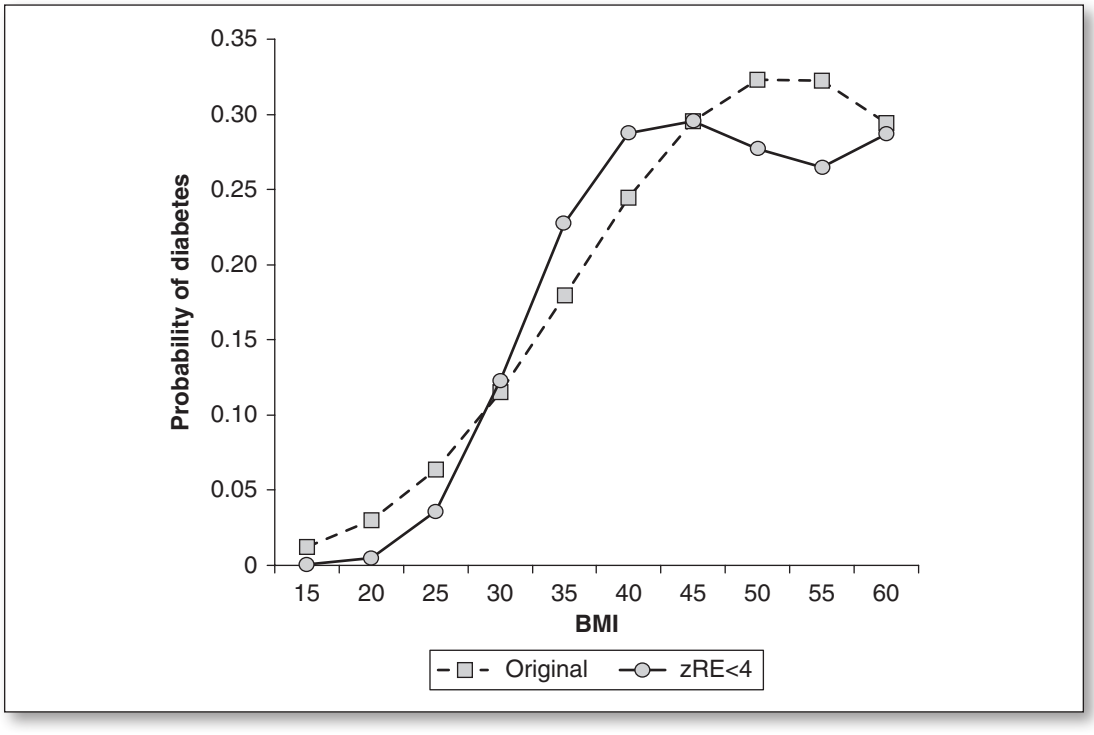
Enrichment Table 7.5 Curvilinear Effects of BMI on Diabetes Following Data Cleaning

Omnibus Tests of Model Coefficients				
		Chi-square	*df*	Sig.
Model 1	Step	1499.272	1	.000
	Block	1499.272	1	.000
	Model	1499.272	1	.000
Model 2	Step	415.846	1	.000
	Block	415.846	1	.000
	Model	1915.118	2	.000
Model 3	Step	108.072	1	.000
	Block	108.072	1	.000
	Model	2023.190	3	.000

Variables in the Equation									
		B	*SE*	Wald	*df*	Sig.	Exp(B)	95% CI for Exp(B)	
								Lower	Upper
Model 1	BMI	.119	.003	1462.817	1	.000	1.126	1.119	1.133
	Constant	–5.922	.101	3453.559	1	.000	.003		
Model 2	BMI	.554	.026	465.272	1	.000	1.741	1.655	1.830
	BMI2	–.006	.000	285.639	1	.000	.994	.993	.995
	Constant	–13.276	.444	893.151	1	.000	.000		
Model 3	BMI	1.4446	.097	223.232	1	.000	4.240	3.508	5.125
	BMI2	–0.0298	.002	146.228	1	.000	.971	.966	.975
	BMI3	.000202	.000	99.033	1	.000	1.000	1.000	1.000
	Constant	–23.937	1.218	385.905	1	.000	.000		

Data Source: NHIS2010, Centers for Disease Control and Prevention.

Enrichment Figure 7.4



Data Source: NHIS2010, Centers for Disease Control and Prevention.

Again, the simple cleaning of data provides a clearer picture of the results.

Cleaning the Data Using DIFCHISQ

Performing the same analysis in SAS, requesting DIFCHISQ produces similar, predictable results. The DIFCHISQ results range up to over 98, which is very high for a single case. The 95th percentile is 7.01, and the 99th percentile is 17.75, so let us start with 14 as a reasonably conservative cut-off point for cleaning the data. This removed 366 cases, again a relatively small number given the overall size of the sample.

For the sake of succinctness, I entered all three terms on the same step in SAS as we knew the results would be significant for all three terms:

Enrichment Table 7.6 Same Analysis With DIFCHISQ Data Cleaning

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	14803.834	12595.777
SC	14812.012	12628.488
–2 Log L	14801.834	12587.777

Data Source: NHIS2010, Centers for Disease Control and Prevention.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	*df*	Pr > ChiSq
Likelihood Ratio	2214.0575	3	<.0001
Score	2083.5613	3	<.0001
Wald	1410.3200	3	<.0001

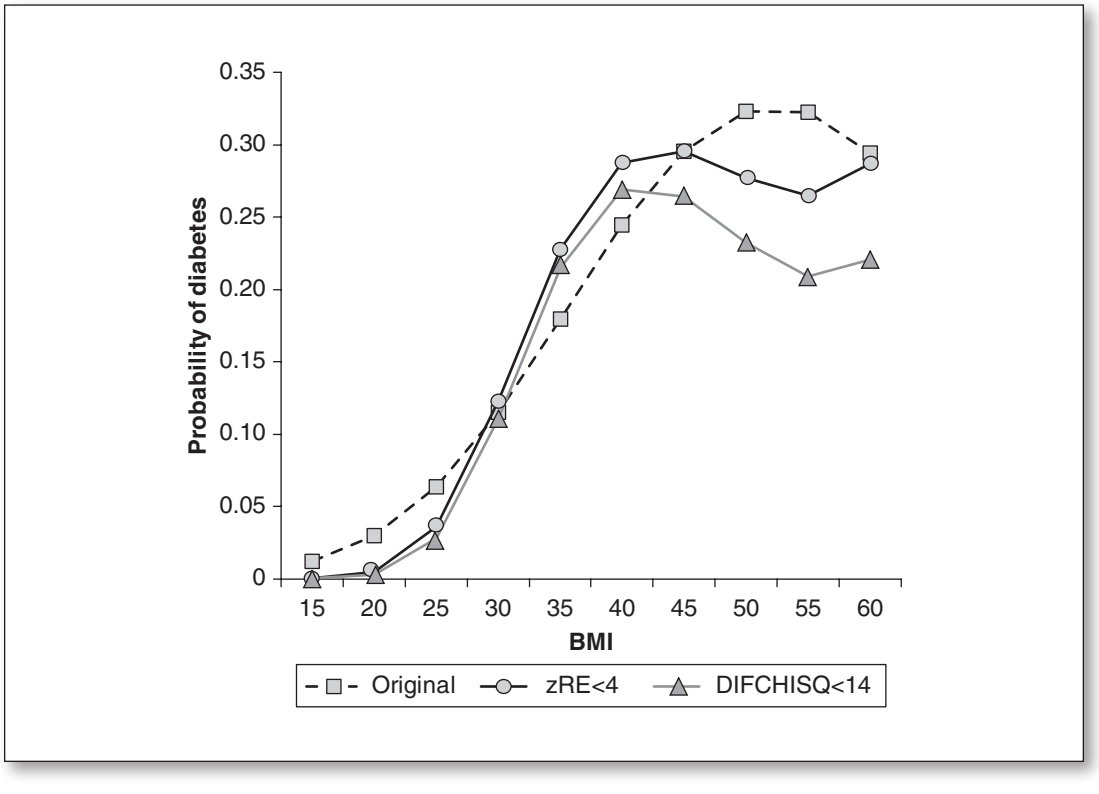
Data Source: NHIS2010, Centers for Disease Control and Prevention.

Analysis of Maximum Likelihood Estimates					
Parameter	*df*	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	–27.1144	1.3341	413.0726	<.0001
BMI	1	1.6609	0.1051	249.9350	<.0001
BMI2	1	–0.0346	0.00267	168.1633	<.0001
BMI3	1	0.000235	0.000022	116.0424	<.0001

Data Source: NHIS2010, Centers for Disease Control and Prevention.

This data cleaning produces a slightly different curve from the others:

Enrichment Figure 7.5



Data Source: NHIS2010, Centers for Disease Control and Prevention.

4. SES and Retention

In this example, I used two different indicators for data cleaning—standardized residuals and DIFCHISQ. Below is an example of what you might see. First, the original analysis (Enrichment Table 7.7).

Enrichment Table 7.7 SES and Retention Curvilinear Analysis Without Data Cleaning

Model 1			
Test	Chi-Square	*df*	Pr > ChiSq
Likelihood Ratio	441.0694	1	<.0001
Score	411.9929	1	<.0001
Wald	389.9710	1	<.0001
Model 2			
Test	Chi-Square	*df*	Pr > ChiSq
Likelihood Ratio	449.9023	2	<.0001
Score	433.9340	2	<.0001
Wald	353.6993	2	<.0001
Model 3			
Test	Chi-Square	*df*	Pr > ChiSq
Likelihood Ratio	453.2357	3	<.0001
Score	437.9948	3	<.0001
Wald	365.9897	3	<.0001

Data Source: NHIS2010, Centers for Disease Control and Prevention.

Model 1					
Parameter	*df*	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.7841	0.0378	5413.6761	<.0001
zBYSES	1	0.7217	0.0365	389.9710	<.0001
Model 2					
Intercept	1	2.7216	0.0433	3949.2397	<.0001
zBYSES	1	0.7909	0.0463	291.9829	<.0001
zBYSES2	1	0.0973	0.0338	8.2752	0.0040
Model 3					
Parameter	*df*	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.7582	0.0480	3303.3788	<.0001
zBYSES	1	0.8809	0.0677	169.4428	<.0001
zBYSES2	1	0.0615	0.0369	2.7786	0.0955
zBYSES3	1	–0.0471	0.0255	3.4259	0.0642

Data Source: NHIS2010, Centers for Disease Control and Prevention.

Following modest cleaning, eliminating DIFCHISQ > 10, we have much stronger results (Enrichment Figure 7.6).

Enrichment Table 7.8 Retention and SES Following Data Cleaning Eliminating DIFCHISQ >10

Testing Global Null Hypothesis: BETA = 0			
Test	Chi-Square	*df*	Pr > ChiSq
Likelihood Ratio	1346.7934	3	<.0001
Score	1120.1784	3	<.0001
Wald	218.6174	3	<.0001

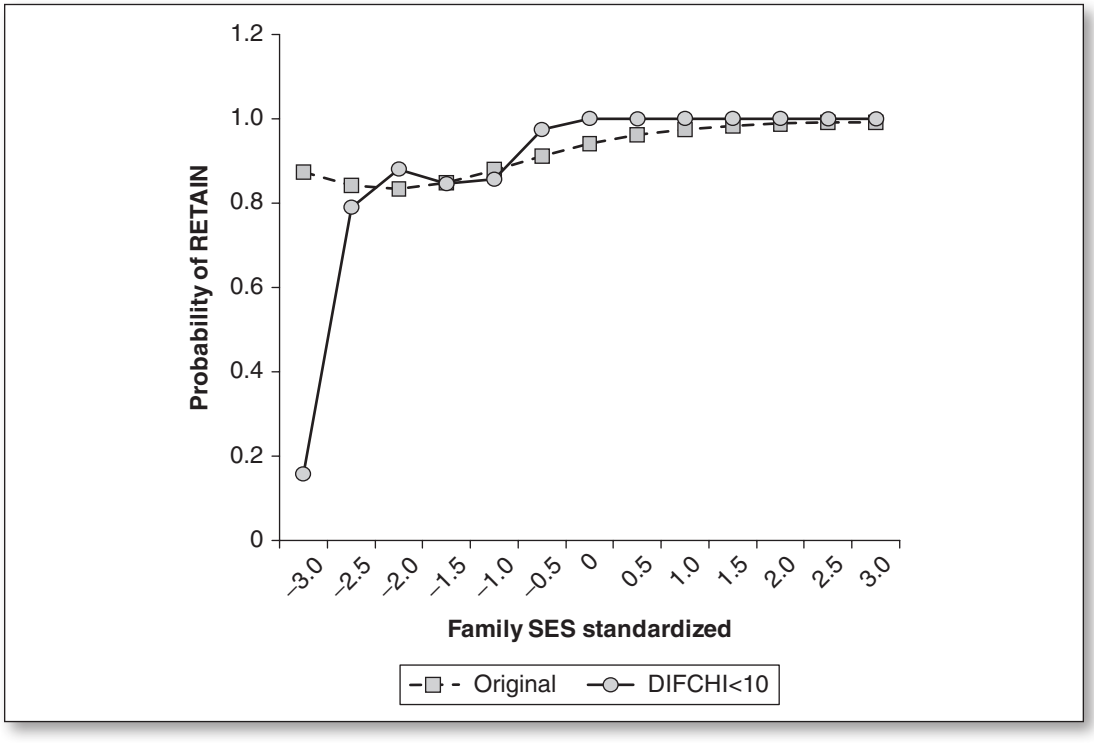
Data Source: NHIS2010, Centers for Disease Control and Prevention.

Analysis of Maximum Likelihood Estimates					
Parameter	*df*	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	8.5346	0.5146	275.0821	<.0001
zBYSES	1	13.8257	1.2869	115.4211	<.0001
zBYSES2	1	8.8874	0.9990	79.1443	<.0001
zBYSES3	1	1.8043	0.2397	56.6467	<.0001

Data Source: NHIS2010, Centers for Disease Control and Prevention.

Thus, we begin with an effect that was only significant in a linear sense and end with some rather strong curvilinear effects by modest data cleaning.

Enrichment Figure 7.6



Data Source: NHIS2010, Centers for Disease Control and Prevention.

SYNTAX EXAMPLES

Example SPSS Syntax to create squared and cubed terms for exploring curvilinearity:

```
compute zBYACH2=zBYACH**2.
compute zBYACH3=zBYACH**3.
execute.
```

Example SPSS Syntax to perform logistic regression entering squared and cubed terms on separate steps:

```
LOGISTIC REGRESSION VARIABLES EVER_MJ
/METHOD=ENTER zBYACH
/METHOD=ENTER zBYACH2
/METHOD=ENTER zBYACH3
/SAVE=DFBETA ZRESID
/PRINT=CI(95)
/CRITERIA=PIN(0.05) POUT(0.10) ITERATE(20) CUT(0.5).
```

Example SAS syntax to enter variables one at a time on separate steps:

```
PROC LOGISTIC DATA=book.ELS2002 descending ;
MODEL GRADUATE = ZBYSES;
;
run;
PROC LOGISTIC DATA=book.ELS2002 descending ;
MODEL GRADUATE = ZBYSES ZBYSES2
;
run;
PROC LOGISTIC DATA=book.ELS2002 descending ;
MODEL GRADUATE = ZBYSES ZBYSES2 ZBYSES3
/selection=none sequential;
run;
```

The above syntax might not be the most elegant, but it allows for direct comparison of separate models by comparing change in −2LL and regression equation at each step. There are options if you use stepwise entry methods to accomplish this with one command (such as the "sequential" command), but stepwise entry methods have the potential to produce problematic outcomes under certain circumstances (like a regression model that includes zBYSES and $zBYSES^3$ but not $zBYSES^2$). Thus, I prefer the above method that provides absolute control to the analyst.

REFERENCES

Aiken, L. S., & West, S. (1991). *Multiple regression: Testing and interpreting interactions* Thousand Oaks, CA: Sage.

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