regression to simply predict Y for a given set of data), we could come to quite different conclusions for the two data sets. Darlington (1968) presents an interesting discussion of this issue and concludes that β_i has only limited utility as a measure of "importance." An even stronger stand is taken by Cooley and Lohnes (1971), who point out that our estimate of β ultimately relies on our estimates of the elements of the intercorrelation matrix. Because this matrix contains $p + p(p-1)/2$ intercorrelations that are all subject to sampling error, Cooley and Lohnes suggested that we must be exceedingly careful about attaching practical significance to the regression coefficients.

As an illustration of the variability of the regression coefficients, a second set of 50 courses, the same set as used for cross-validation, was drawn from the same source as that for the data in Table 15.1. In this case, R^2 was more or less the same as it had been for the first example ($R^2 = .710$), but the regression equation looked quite different. In terms of standardized variables,

$$Z_{\hat{Y}} = 0.371\,\text{Teach} + 0.113\,\text{Exam} + 0.567\,\text{Knowledge} - 0.27\,\text{Grade} + 0.184\,\text{Enroll}$$

If you compare this equation with the one found from Exhibit 15.1, it is clear that there are substantial differences in some of the values of β_i.

Another measure of importance, which has much to recommend it, is the squared semipartial correlation between predictor i and the criterion (with all other predictors partialled out)—that is, $r^2_{0(i.123\ldots p)}$. Darlington (1968) refers to this measure as the "usefulness" of a predictor. As we have already seen, this semipartial correlation squared represents the decrement in R^2 that would result from the elimination of the ith predictor from the model (or the increment that would result from its addition). When the main goal is prediction rather than explanation, this is probably the best measure of "importance." Fortunately, it is easy to obtain from most computer printouts, because

$$r^2_{0(i.123\ldots p)} = \frac{F_i(1 - R^2_{0.123\ldots p})}{N - p - 1}$$

where F_i is the F test on the individual β_i (or b_i) coefficients. (If your program uses t tests on the coefficient, $F = t^2$.) Because all terms except F_i are constant for $i = 1 \ldots p$, the F_is order the variables in the same way as do the squared semipartials and, thus, can be used to rank order the variables in terms of their usefulness.

Darlington (1990) has made a strong case for not squaring the semipartial correlation when speaking about the importance of variables. His case is an interesting one. However, whether or not the correlations are squared will not affect the ordering of variables. (If you want to argue persuasively about the absolute importance of a variable, you should read Darlington's argument.)

One common, but unacceptable, method of ordering the importance of variables is to rank them by the order of their inclusion in a stepwise regression solution. The problem with this approach is that it ignores the interrelationships among the variables. Thus, the first variable to be entered is entered solely on the strength of its correlation with the criterion. The second variable entered is chosen on the basis of its correlation with the criterion after partialling the first variable but ignoring all others. The third is chosen on the basis of how it correlates with the criterion after partialling the first two variables, and so on. In other words, each variable is chosen on a different basis, and it makes little sense to rank them according to order of entry. To take a simple example, assume that variables 1, 2, and 3 correlate .79, .78, and .32 with the criterion. Assume further that variables 1 and 2 are correlated .95, whereas 1 and 3 are correlated .20. They will then enter the equation in the order 1, 3, and 2, with the last entry being nonsignificant. But in what sense do we mean to say that variable 3 ranks above variable 2 in importance? I would hate to defend such a statement to a reviewer—actually, I would be hard pressed even to say what I meant by importance in this situation. A similar point has been made well by Huberty (1989). For an excellent discussion of measures of importance, see Harris (1985, pp. 79ff).

15.12 Using Approximate Regression Coefficients

I have pointed out that regression coefficients frequently show substantial fluctuations from sample to sample without producing drastic changes in R. This might lead someone to suggest that we might use rather crude approximations of these coefficients as a substitute for the more precise estimates obtained from the data. For example, suppose that a five-predictor problem produced the following regression equation:

$$\hat{Y} = 9.2 + 0.85X_1 + 2.1X_2 - 0.74X_3 + 3.6X_4 - 2.4X_5$$

We might ask how much loss we would suffer if we rounded these values to

$$\hat{Y} = 10 + 1X_1 + 2X_2 - 1X_3 + 4X_4 - 2X_5$$

The answer is that we would probably lose very little. Excellent discussions of this problem are given by Cohen et al. (2003), Dawes and Corrigan (1974), and Wainer (1976, 1978).

This method of rounding off regression coefficients is more common than you might suppose. For example, the college admissions officer who quantifies the various predictors he has available and then weights the grade point average twice as highly as the letter of recommendation is really using crude estimates of what he thinks would be the actual regression coefficients. Similarly, many scoring systems for the Minnesota Multiphasic Personality Inventory (MMPI) are based on the reduction of coefficients to convenient integers. Whether the use of these *diagnostic signs* produces results that are better than, worse than, or equivalent to the use of the usual linear regression equations is still a matter of debate. A dated but very comprehensive study of this question is presented in Goldberg (1965). Rather than undermining our confidence in multiple regression, I think the fact that rounded off coefficients do nearly as well (sometimes better if we are applying them to new data) speaks to the robustness of regression. It also suggests that you not put too much faith in small differences in coefficients.

15.13 Mediating and Moderating Relationships

One of the most frequently cited papers in the psychological literature related to multiple regression during the past 20 years has been a paper by Baron and Kenny (1986) on what they called the moderator-mediator distinction. The important point for both moderating and mediating relationships is that a third variable plays an important role in governing the relationship between two other variables.

Mediation

mediating relationship

A **mediating relationship** is what it sounds like—some variable mediates the relationship between two other variables. For example, take a situation in which high levels of care from your parents leads to feelings of competence and self-esteem on your part, which, in turn, leads to high confidence when you become a mother. Here, we would say that your feelings of competence and self-esteem *mediate* the relationship between how you were parented and how you feel about mothering your own children.

Baron and Kenny (1986) laid out several requirements that must be met before we can speak of a mediating relationship. Consider the diagram in Figure 15.5 as being representative of a mediating relationship that we want to explain.

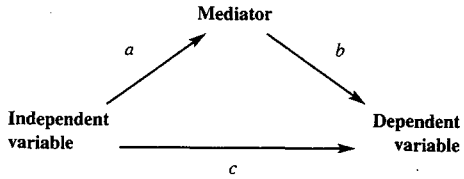

Figure 15.5 Diagram of a mediated relationship

The predominant relationship that we want to explain is labeled "c" and is the path from the independent to the dependent variable. The mediating path has two parts: "a," the path connecting the independent variable to the potential mediator, and "b," the path connecting that mediator to the dependent variable.

Baron and Kenny argued that for us to claim a mediating relationship, we need to first show that there is a significant relationship between the independent variable and the mediator. (If the mediator is not associated with the independent variable, then it couldn't mediate anything.) The next step is to show that there is a significant relationship between the mediator and the dependent variable, for reasons similar to those for the first requirement. Then we need to show that there is a significant relationship between the independent and dependent variable.

These three conditions require that the three paths (a, b, and c) are all individually significant. The final step consists of demonstrating that when the mediator and the independent variable are used simultaneously to predict the dependent variable, the previously significant path between the independent and dependent variables (c) is now greatly reduced, if not nonsignificant. Maximum evidence for mediation would occur if c drops to 0. I have never seen a path go away completely. Most likely to happen is that c becomes a weaker, though perhaps still significant, path.

Leerkes and Crockenberg (1999) were interested in studying the relationship between how children were raised by their own mothers, and their later feelings of maternal self-efficacy when they, in turn, became mothers. The sample consisted on 92 mothers of 5-month old infants. The researchers expected to find that high levels of maternal care when the mother was a child translated to high levels of self-efficacy when that child later became a mother. But Leerkes and Crockenberg went further, postulating that the mediating variable in this relationship is self-esteem. They argued that high levels of maternal care lead to high levels of self-esteem in the child and that this high self-esteem later translates into high levels of self-efficacy as a mother. This relationship is diagrammed in Figure 15.6.

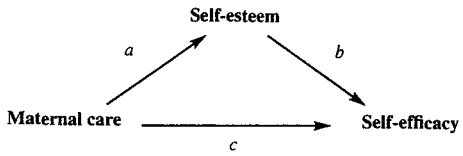

Figure 15.6 Diagram of mediation in the development of self-efficacy

The initial conditions of Baron and Kenny (1986) can be tested by looking at the simple correlations among the variables. These are shown here, as produced by SPSS.

Correlations

Pearson Correlation

	Maternal care	Self-esteem	5 month efficacy
Maternal care	1.000	.403**	.272**
Self-esteem	.403**	1.000	.380**
5 month efficacy	.272**	.380**	1.000

** Correlation is significant at the 0.01 level (2-tailed).

Exhibit 15.4a Correlations between variables in mediation example

Here, we can see that maternal care is correlated with self-esteem and with self-efficacy, and that self-esteem is also correlated with self-efficacy. These relationships satisfy Baron and Kenny's basic prerequisites. The next step is to use both self-esteem and maternal care as predictors of self-efficacy. This is shown in Exhibit 15.4b.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations	
		B	Std. Error	Beta	t	Sig.	Zero-order	Part
1	(Constant)	3.260	.141		23.199	.000		
	maternal care	.112	.042	.272	2.677	.009	.272	.272
2	(Constant)	2.929	.173		16.918	.000		
	maternal care	5.817E-02	.044	.142	1.334	.185	.272	.130
	self esteem	.147	.048	.323	3.041	.003	.380	.295

[a] Dependent Variable: 5 month efficacy

Exhibit 15.4b Multiple regressions for mediation example

The first model in the previous table uses maternal care as the sole predictor. The second model has added self-esteem as a predictor. You can see that when we add self-esteem to maternal care, which was clearly significant when used alone to predict self-efficacy, maternal care is no longer significant ($t = 1.334$, $p = 0.185$). This is evidence that self-esteem is serving a mediating role between maternal care and self-efficacy. The output also shows what SPSS calls the "part correlation," but which the rest of us call the semipartial correlation. The semipartial correlation between maternal care and self-efficacy is .130, whereas the simple correlation (zero-order) between maternal care and self-efficacy was .27. It remains significant, as we can see by the t test on self-esteem, but has dropped noticeably.

These results support Leerkes and Crockenberg's hypothesis that self-esteem played a mediating role between maternal care and self-efficacy. Caring parents seem to produce children with higher levels of self-esteem, and this higher self-esteem translates into positive feelings of self-efficacy when the child, in turn, becomes a mother.

In this situation, Leerkes and Crockenberg were fortunate to have a situation in which the direct path from maternal care to self-efficacy dropped to nonsignificance when self-esteem was added. Unfortunately, that does not always happen. (Actually, it seems to happen relatively infrequently.) The more common result is that the direct path becomes less important, though it remains significant. There has been considerable discussion about what to do in this situation, but there is a relatively simple answer, developed by Sobel (1982), that was referred to by Baron and Kenny.

When we have a situation in which the direct path remains significant, though at a lower value, one way to test for a mediating relationship is to ask whether the complete mediating path from independent variable to mediator to dependent variable is significant. To do this, we need to know the regression coefficients and their standard errors for the two paths in the mediating chain. We will soon also need the regression of Self-esteem on Maternal care, so that table follows.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations		
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part
1	(Constant)	2.257	.294		7.687	.000			
	pbi maternal care	.364	.087	.403	4.178	.000	.403	.403	.403

[a] Dependent Variable: self esteem

Exhibit 15.4c Regression of self-esteem on maternal care

The important statistics from the two regressions are shown in Table 15.6. Because SPSS does not report the standard error of beta, we need to calculate it. The t statistic given in these tables is either the unstandardized regression coefficient (b) divided by its standard error, or the standardized regression coefficient divided by its standard error. Thus, we can solve

$$t = \frac{\beta}{s_\beta}; \qquad s_\beta = \frac{\beta}{t} = \frac{0.403}{4.178} = 0.096$$

Similarly for the path from Self-esteem to Self-efficacy, partialling Maternal care, we have

$$t = \frac{\beta}{s_\beta}; \qquad s_\beta = \frac{\beta}{t} = \frac{0.323}{3.041} = 0.106$$

These results yield Table 15.6.

Table 15.6 Regression coefficients and standard errors for two parts of mediating path

Path *a*		Path *b*	
Maternal Care →	Self-Esteem	Self-Esteem →	Self-Efficacy
β_a	0.403	β_b	0.380
s_a	0.096	s_b	0.098
t	4.18*	t	3.89*

Then the regression coefficient for the path from Maternal care → Self-esteem → Self-efficacy is equal to $\beta_a \times \beta_b = 0.403 \times 0.323 = 0.130$, where a and b refer to the relevant paths. (Path c is the direct path from Maternal care to Self-efficacy.) In addition, we know that the standard error of this two-part path is given by

$$s_{\beta_a\beta_b} = \sqrt{\beta_a^2 s_b^2 + \beta_b^2 s_a^2 - s_a^2 s_b^2}$$

where β_a and β_b are the paths, and s_a^2 and s_b^2 are the corresponding standard errors of the standardized regression coefficients for those paths.[14] We can calculate the standard error of the combined path as

$$\begin{aligned} s_{\beta_a\beta_b} &= \sqrt{\beta_a^2 s_b^2 + \beta_b^2 s_a^2 - s_a^2 s_b^2} = \sqrt{.403^2(.106^2) + .323^2(.096^2) - (.106^2)(.096^2)} \\ &= \sqrt{0.0027} \\ &= 0.052 \end{aligned}$$

We now know the path coefficient ($0.403 \times 0.323 = 0.130$) and its standard error (0.052), and we can form a t ratio as

$$t = \frac{\beta_1\beta_3}{s_{\beta_1\beta_3}} = \frac{.130}{.052} = 2.50$$

Sobel (1982) stated that this ratio is asymptotically normally distributed, which, for large samples, would lead to rejection of the null hypothesis at $\alpha = 0.05$ when the ratio exceeds ± 1.96. It would presumably have a t distribution on $N - 3$ df for small samples. In our case, the path is clearly significant, as we would expect from the previous results. Therefore, we can conclude that we have convincing evidence of a strong mediating pathway from maternal care through self-esteem to self-efficacy. Because the regression coefficient (and semi-partial correlation) for the direct path from maternal care to self-efficacy is not significant, the main influence of maternal care is through its mediating relationship with self-esteem.

There has been considerable discussion in the literature about the best approach to testing mediation. For an online test using three alternative approaches to the standard error, go to www.unc.edu/~preacher/sobel/sobel.htm. Preacher and Hayes (2004) (available from the previous website) present SPSS and SAS macros that allow you to using bootstrapping methods (see Chapter 18) to address this question. A very well-written description of mediation has been put on the Web by Paul Jose, at the University of Wellington. It can be found at http://www.vuw.ac.nz/psyc/staff/paul-jose/files/helpcentre/help7_mediation_example.php. In additon, Jose offers a free mediation calculator, which runs under Excel, at http://www.vuw.ac.nz/psyc/staff/paul-jose/files/medgraph/medgraph.php. I have found that very useful, but be aware that there seems to be minor disagreement between the example and the results of the software. Finally, an extensive comparison of alternative approaches can be found in MacKinnon, Lockwood, Hoffman, West, and Sheets (2002).

Moderating Relationships

moderating relationships

Whereas a mediating relationship attempts to identify a variable or variables through which the independent variable acts to influence the dependent variable, **moderating relationships** refer to situations in which the relationship between the independent and dependent variables changes as a function of the level of a third variable (the moderator).

[14] There is some disagreement about the exact form of these equation, but the one given here is recommended by Baron and Kenny. The differences between the various equations turn out to be very minor in practice.

Wagner, Compas, and Howell (1988) hypothesized that individuals who experience more stress, as assessed by a measure of daily hassles, will exhibit higher levels of symptoms than will those who experience little stress. That is what, in analysis of variance terms, would be the main effect of hassles. However, the researchers also expected that if a person had a high level of social support to help deal with his or her stress, symptoms would increase only slowly with increases in hassles. For those who had relatively little social support, symptoms were expected to rise more quickly as hassles increased.

Wagner et al. (1988) studied students who were attending an orientation before starting their first year of collge. Students were asked to report on the number of minor stressful events (labeled hassles) that they had recently experienced and to report on their perceived level of social support. Students then completed a symptom checklist about the number of symptoms they had experienced in the past month. For this part of the study, there were complete data on 56 participants. These data are available in a file named hassles.dat (from www.uvm.edu/~dhowell/methods/).

Our first step is to look at the relationships between these variables. The correlation matrix is shown here.

Correlations

Pearson Correlation

	Hassles	Support	Symptoms
Hassles	1.000	−.167	.577**
Support	−.167	1.000	−.134
Symptoms	.577**	−.134	1.000

** Correlation is significant at the 0.01 level

As expected, there is a significant relationship between Hassles and Symptoms ($r = .577$), though Support is not related to Symptoms, or to Hassles. This does not, however, answer the question that the researchers really wanted to ask, which is whether the relationship between Hassles and Symptoms depends on the degree of social support.

If you think about this question, it starts to sound very much like the question behind an interaction in the analysis of variance. Actually, it is an interaction, and the way that we will test for that interaction is to create a variable that is the product of Hassles and Support. (This is also similar to what we will do in the general linear model approach to the analysis of variance in the next chapter.) However, if we just multiply Hassles and Support together, there will be two problems with what results. In the first place, either Hassles or Support or both will be highly correlated with their product, which will make for multicollinearity in the data. This will seriously effect the magnitude, and tests of significance, of the coefficients for the main effect of Hassles and Support. The second problem is that any effect of Hassles or Support in the regression analysis will be evaluated at a value of 0 for the other variable. In other words, the test on Hassles will be a test on whether Hassles is related to Symptoms if a participant had exactly no social support. Similarly, the test on Support would be evaluated for those participants who have exactly no hassles. Both the problem of multicollinearity and the problem of evaluating one main effect at an extreme value of the other main effect are unwelcome.

center

To circumvent these two problems, we are going to **center** our data. This means that we are going to create deviation scores by subtracting each variable's mean from the individual observations. Now a score of 0 for (centered) Hassles represents someone who has the mean level of Hassles, which seems an appropriate place to examine any effects of support, and anyone with a 0 on (centered) support represents someone with a mean level of support. This has solved one of our problems because we are now evaluating the main effects at a reasonable

level of the other main effect. It has also helped to solve our other problem because if you look at the resulting correlations, multicollinearity will have been significantly reduced.

Having centered our variables, we will then form a product of our centered variables, and this will represent our interaction term. The means for hassles, support, and symptoms are 170.1964, 28.9643, and 90.4286, respectively, and the equations for creating centered variables and their interaction follow. The letter "c" at the beginning of the variable name indicates that it is centered.

$$\text{chassles} = \text{hassles} - 170.1964$$

$$\text{csupport} = \text{support} - 28.9643$$

$$\text{chassupp} = \text{chassles} \times \text{csupport}$$

The correlations among the centered (and uncentered) variables are shown in the following table. I have included the product of the uncentered variables simply to show how high the correlation between hassles and hassupp is, but we are not going to use this variable. You can see that by centering the variables we have substantially reduced the correlation between the main effects and the interactions. That was our goal. Notice that centering the variables did not change their correlations with each other—only with the interaction.

Correlations

Pearson Correlation

	Hassles	Support	Symptoms	hassupp	chassles	csupport	chassupp
Hassles	1.000	−.167	.577**	.910**	1.000**	−.167	−.297*
Support	−.167	1.000	−.134	−.510**	−.167	1.000**	.402**
Symptoms	.577**	−.134	1.000	.585**	.577**	−.134	−.391**
hassupp	.910**	−.510**	.585**	1.000	.910**	−.510**	−.576**
chassles	1.000**	−.167	.577**	.910**	1.000	−.167	−.297*
csupport	−.167	1.000**	−.134	−.510**	−.167	1.000	.402**
chassupp	−.297*	.402**	−.391**	−.576**	−.297*	.402**	1.000

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

We can now examine the interaction of the two predictor variables by including the interaction term in the regression with the other centered predictors. The dependent variable is Symptoms. This regression is shown in Exhibit 15.5. (As long as we use the product of centered variables, it doesn't matter, except for the intercept, if we use the centered or uncentered main effects. I prefer the latter, but for no particularly good reason.)

From the printout, you can see that $R^2 = .388$, which is significant. (Without the interaction term, R^2 would have been .334 [not shown].) From the table of regression coefficients, you see that both the centered Hassles and the interaction terms are significant ($p = .000$ and .037, respectively), but the social support variable is not significant. By convention, we leave it in our regression solution because it is involved in the interaction, even though the associated t value shows that deleting that variable would not lead to a significant decrease in R^2.

Our regression equation now becomes

$$\hat{Y} = .086\text{chassles} + 0.146\text{csupport} - .005\text{chassupp} + 89.585.$$

We have answered our initial questions (social support does moderate the relationship between hassles and symptoms), but it would be helpful if we could view this graphically to

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.623[a]	.388	.353	16.8932

[a] Predictors: (Constant), CHASSUPP, CHASSLES, CSUPPORT

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9427.898	3	3142.633	11.012	.000[a]
	Residual	14839.816	52	285.381		
	Total	24267.714	55			

[a] Predictors: (Constant), chassupp, chassles, csupport

[b] Dependent Variable: Symptoms

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	89.585	2.292		39.094	.000
	chassles	8.594E-02	.019	.509	4.473	.000
	csupport	.146	.305	.057	.479	.634
	chassupp	-5.06E-03	.002	-.262	-2.144	.037

[a] Dependent Variable: Symptoms

Exhibit 15.5 Regression solution for moderated relationship between hassles and symptoms

interpret the meaning of the interactive effect. Excellent discussions of this approach can be found in Finney, Mitchell, Cronkite, and Moos (1984), Jaccard, Turrisi, and Wan (1990), and Aiken and West (1991). The latter is the authoritative work on moderation. Normand Péladeau has a free program called Italassi, available on the web at http://www.simstat.com/. This program will plot the interaction on your screen and provides a slider so that you can vary the level of the support variable.

The simplest solution is to look at the relationship between chassles and csymptoms for fixed levels of social support. Examination of the distribution of csupport scores shows that they range from about -21 to $+19$. Thus, scores of -15, 0, and $+15$ would represent low, neutral, and high scores on csupport. (You don't have to be satisfied with these particular values, you can use any that you like. I have picked extremes to better illustrate what is going on.

First, I will rewrite the regression equation, substituting generic labels for the regression coefficients. I will also substitute chassles × csupport for chassupp because that is the way that I calculated chssupp. Finally, I will also reorder the terms a bit just to make life easier.

$$\hat{Y} = b_1\text{chassles} + b_2\text{csupport} - b_3\text{chassupp} + b_0$$

$$\hat{Y} = b_0 + b_2\text{csupport} + b_3(\text{chassles} \times \text{csupport}) + b_1\text{chassles}$$

Collecting terms, I have

$$\hat{Y} = b_0 + b_2\text{csupport} + \text{chassles}(b_3\text{csupport} + b_1)$$

Next, I will substitute the actual regression coefficients to get

$$\hat{Y} = [89.585 + 0.146\text{csupport}] + \text{chassles}(-.005\text{csupport} + .086)$$

Notice the first term in square brackets. For any specific level of csupport (e.g., 15), this is a constant. Similarly, for the terms in parentheses after chassles, that is also a constant for a fixed level of support. To see this most easily, we can solve for $\hat{Y}$ when csupport is at 15, which is a high level of support. This gives us

$$\hat{Y} = [89.585 + 0.146 \times 15] + \text{chassles}(-.005 \times 15 + .086)$$
$$= 91.755 + 0.011 \times \text{chassles}$$

which is just a plain old linear equation. This is the equation that represents the relationship between $\hat{Y}$ and chassles when social support is high (i.e., 15).

Now we can derive two more simple linear equations, one by substituting 0 for csupport and one by substituting -15.

When csupport = 0,

$$\hat{Y} = 89.585 + .086 \times \text{chassles}$$

When csupport = -15,

$$\hat{Y} = 87.395 + .161 \times \text{chassles}$$

When I look at the frequency distribution of chassles, low, neutral, and high scores are roughly represented by -150, 0, and 150. So I will next calculate predicted values for symptoms and low, neutral, and high levels of chassles for each of low, neutral, and high levels of csupport. These are shown in Table 15.7, and they were computed using the three previous regression equations and setting chassles at -150, 0, and 150.

Table 15.7 Predicted values of symptoms at varying levels of hassles & support

		Centered Support		
		−15	0	15
Centered Hassles	−150	63.245	76.685	90.105
	0	87.395	89.585	91.755
	150	111.545	102.485	93.405

If we plot these predicted values separately for the different levels of social support, we see that with high social support increases in hassles are associated with relatively small increases in symptoms. When we move to csupport = 0, which puts us at the mean level of support, increasing hassles leads to a greater increase in symptoms. Finally, when we have low levels of support (csupport = -15), increases in hassles lead to dramatic increases in symptoms. This is shown graphically in Figure 15.7.

The use of interaction terms (e.g., $X_1 \times X_2$) in data analysis, such as the problem that we have just addressed, has become common in psychology in recent years. However, my experience and that of others has been that it is surprisingly difficult to find meaningful situations where the regression coefficient for $X_1 \times X_2$ is significant, especially in experimental settings where we deliberately vary the levels of X_1 and X_2. McClelland and Judd (1993)

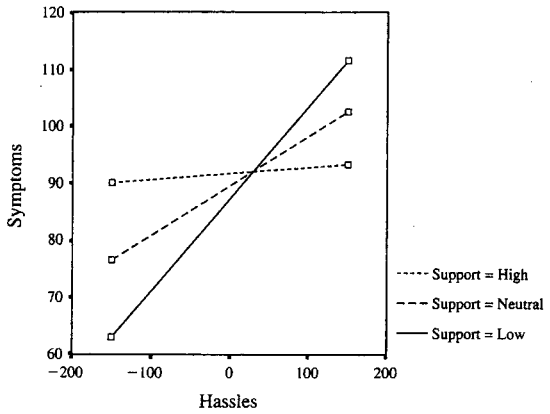

Figure 15.7 Plot of symptoms as a function of hassles for different levels of social support

have investigated this problem and have shown why our standard field study designs have so little power to detect interactions. That is an important paper for anyone investigating interaction effects in nonexperimental research.

15.14 Logistic Regression

logistic regression

In the past few years, the technique of **logistic regression** has become popular in the psychological literature. (It has been popular in the medical and epidemiological literature for much longer.) Logistic regression is a technique for fitting a regression surface to data in which the dependent variable is a dichotomy.[15] A very common situation in medicine is the case in which we want to predict response to treatment, where we might code survivors as 1 and those who don't survive as 0. In psychology, we might class clients as Improved or Not Improved, or we might rate performance as Successful or Not Successful. Whenever we have such a dichotomous outcome, we have a possible candidate for logistic regression.

discriminant analysis

But when we have a dichotomous dependent variable, we have at least two other statistical procedures as candidates for our analysis. One of them, which is not discussed in this text, is **discriminant analysis,** which is a technique for distinguishing two or more groups on the basis of a set of variables. The question is often raised about whether logistic regression is better than discriminant analysis. It isn't always clear how we might define "better," but discriminant analysis has two strikes against it that logistic regression does not. In the first place, discriminant analysis can easily produce a probability of success that lies outside the range of 0 and 1, yet we know that such probabilities are impossible. In the second place, discriminant analysis depends on certain restrictive normality assumptions on the independent variables, which are often not realistic. Logistic regression, on the other hand, does not

[15] Logistic regression can also be applied in situations where there are three or more levels of the dependent variable, which we refer to as a polychotomy, but we will not discuss that method here.

produce probabilities beyond 0 and 1, and requires no such restrictive assumptions on the independent variables, which can be categorical or continuous. Common practice has now moved away from discriminant analysis in favor of logistic regression.

A second alternative would be to run a standard multiple regression solution, which we have just been covering, using the dichotomous variable as our dependent variable. In many situations the results would be very similar. But there are reasons to prefer logistic regression in general, though to explain those I have to use a simple example.

We will look at actual, though slightly modified, data on variables that we hope to relate to whether or not the individual responds positively to cancer treatment. The data that we will consider were part of a study of behavioral variables and stress in people recently diagnosed with cancer. For our purposes, we will look at patients who have been in the study for at least a year, and our dependent variable (Outcome) is coded 1 for those who have improved or are in complete remission, and 0 for those who have not improved or who have died. (Any consistent method of coding, such as 1 and 2, or 5 and 8, would also work.)[16] Out of 66 cases, we have 48 patients who have improved and 18 who have not. Suppose that we start our discussion with a single predictor variable, which is the Survival rating (SurvRate) assigned by the patient's physician at the time of diagnosis. This is a number between 0 and 100 and represents the estimated probability of survival at 5 years.

One way to look at the relationship between SurvRate and Outcome would be to simply create a scatterplot of the two variables, with Outcome on the Y axis. Such a plot is given in Figure 15.8. (In this figure, I have offset overlapping points slightly so that you could see them pile up. That explains why there seems to be string of points at SurvRate = 91 and Outcome = 1, for example.) From this plot, we can see that the proportion of people who

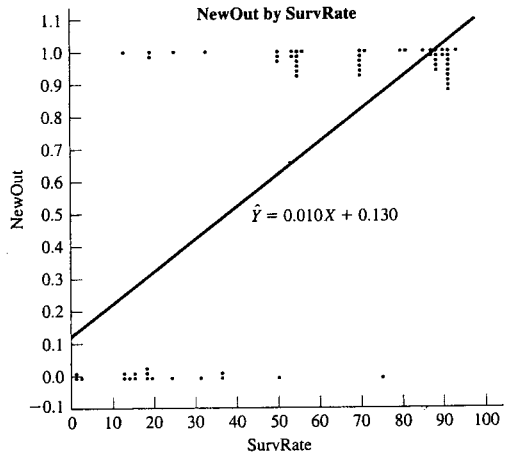

Figure 15.8 Outcome as a function of SurvRate

[16] You have to be careful with coding because different programs treat the same codes differently. Some will code the higher value as success and the lower as failure, and others will do the opposite. If you have a printout where the results seem exactly the opposite of what you might expect, check the manual to see how the program treats the dichotomous variable.

improve is much higher when the survival rating is high, as we would expect. Assume for the moment that we had a great many subjects and could calculate the mean Outcome score (the mean of 0s and 1s) associated with each value of SurvRate. (These are called **conditional means** because they are conditional on the value of SurvRate.) The conditional means would be the proportion of people with that value of SurvRate who improved. If we fit a standard regression line to these data, this would be the regression line that fits the *probability* of improvement as a function of SurvRate. But as you can imagine, for many values of SurvRate, the predicted probability would be outside the bounds 0 and 1, which is impossible. That alone would make standard linear regression a poor choice. There is a second problem. If you were to calculate the *variances* of Outcome for different values of SurvRate, you would see that they are quite small for both large and small values of SurvRate (because almost everyone with low values of SurvRate has a 0 and almost everyone with high values of SurvRate has a 1). But for people with mid-level SurvRate values, there is nearly an even mix of 0s and 1s, which will produce a relatively larger variance. This will clearly violate our assumption of homogeneity of variance in arrays, to say nothing of normality. Because of these problems, standard linear regression is not a wise choice with a dichotomous dependent variable, though it would provide a pretty good estimate if the percentage of improvement scores didn't fall below 20% or above 80% across all values of SurvRate (Cox & Wermuth, 1992).

conditional means

Another problem is that the true relationship is not likely to be linear. Differences in SurvRate near the center of the scale will lead to noticeably larger differences in Outcome than will comparable differences at the ends of the scale.

sigmoidal

Although a straight line won't fit the data in Figure 15.6 well, an S-shaped, or **sigmoidal** curve will. This line changes little as we move across low values of SurvRate, then changes rapidly as we move across middle values, and finally changes slowly again across high values. In no case, does it fall below 0 or above 1. This line is shown in Figure 15.9. Notice that it is quite close to the whole cluster of points in the lower left, rises rapidly for those values of SurvRate that have a roughly equal number of patients who improve and don't improve, and then comes close to the cluster of points in the upper right. When you think about how you might expect the probability of improvement to change with SurvRate, this curve makes sense.

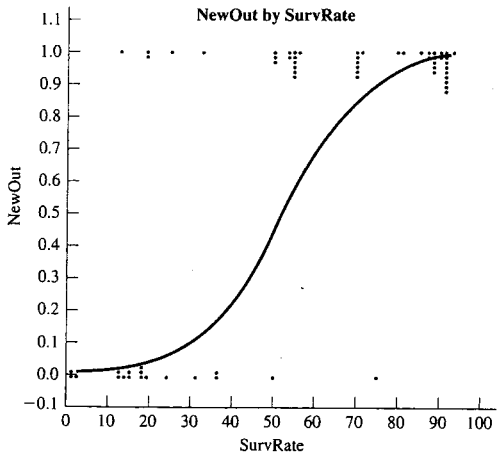

Figure 15.9 More appropriate regression line for predicting outcome

There is another way to view what is happening that provides a tie to standard linear regression. If you think back to what we have said in the past about regression, you will recall that, at least with large samples, a whole collection of Y values correspond to each value of X. You saw this diagrammatically in Figure 9.5, when I spoke about the assumptions of normality and homogeneity of variance in arrays. Rather than classifying people as improved or not improved, suppose that we could somehow measure their disease outcomes more precisely. Then for a rating of SurvRate = 20, for example, we would have a whole distribution of disease outcome scores; similarly for people with SurvRate = 30, SurvRate = 40, and so on. These distributions are shown schematically in Figure 15.10.

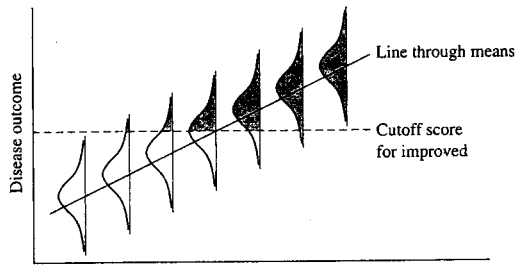

Figure 15.10 Disease outcome as a function of SurvRate

When we class someone as improved, we are simply saying that his disease outcome score is sufficiently high for us to say that he falls in that category. He may be completely cured, he may be doing quite a bit better, or he may be only slightly improved, but he at least met our criterion of "improved." Similarly, someone else may have remained constant, gotten slightly worse, or died, but in any event her outcome was below our decision point.

censored data

What we have here are called **censored data.** When I speak of censoring, I'm not talking about some nasty little man with a big black marker who blocks out things he doesn't want others to see. We are talking about a situation where something that is above a cutoff is classed as a success, and something below the cutoff is classed as a failure. It could be performance on a test, obtaining a qualifying time for the Boston Marathon, or classifying an airline flight as "on time" or "late." From this point of view, logistic regression can be thought of as applying linear regression to censored data. Because the data are censored to provide only success or failure, we have to fit our model somewhat differently.

The horizontal line across the plot in Figure 15.8 represents a critical value. Anyone scoring above that line would be classed as improved, and anyone below it would be classed as not improved. As you can see, the proportion improving, as given by the shaded area of each curve, changes slowly at first, then much more rapidly, and then slowly again as we move from left to right. This should remind you of the sigmoid curve we saw in Figure 15.9 because this is what gives rise to that curve. The regression line that you see in Figure 15.10 is the linear regression of the *continuous* measure of outcome against SurvRate, and it goes through the mean of each distribution. If we had the continuous measure, we could solve for this line. But we have censored data, containing only the dichotomous values, and for that we are much better off solving for the sigmoidal function in Figure 15.9.

We have seen that although our hypothetical continuous variable is a linear function of SurvRate, our censored dichotomous variable (or the probability of improvement) is

not. But a simple transformation from p(improvement) to odds(improvement) to log odds(improvement) will give us a variable that *is* a linear function of SurvRate. Therefore, we can convert p(improvement) to log odds(improvement) and get back to a linear function.

Dabbs and Morris (1990) ran an interesting study in which they classified male military personnel as High or Normal in testosterone, and as either having, or not having, a history of delinquency. The results follow:

		Delinquent		
		Yes	No	Total
Testosterone	Normal	402	3614	4016
	High	101	345	446
		503	3959	4462

For these data, the odds of being delinquent if you are in the Normal group are (frequency delinquent)/(frequency not delinquent). (Using probabilities instead of frequencies, this comes down to $p_{\text{delinquent}}/p_{\text{not delinquent}} = p(\text{delinquent})/(1 - p(\text{delinquent}))$.) For the Normal testosterone group, the odds of being delinquent are $402/3614 = .1001$. The odds of being not delinquent if you are in the Normal group is the reciprocal of this, which is $3614/402 = 8.990$. This last statistic can be read as meaning that if you are a male with normal testosterone levels, you are nearly 9 times *less likely* to be delinquent than not delinquent. If we look at the High testosterone group, however, the odds of being delinquent are $101/345 = 0.293$, and the odds of being not delinquent are $345/101 = 3.416$. Both groups of males are more likely to be not delinquent than delinquent, but that isn't saying much, because we would hope that most people are not delinquent. But notice that as you move from the Normal to the High group, your odds of being delinquent nearly triple, going from 0.111 to 0.293. If we form the ratio of these odds, we get $0.293/0.111 = 2.64$, which is the odds ratio. For these data, you are 2.64 times more likely to be delinquent if you have high testosterone levels than if you have normal levels. That is a pretty impressive statistic.

We will set aside the odds ratio for a moment and just look at odds. With our cancer data, we will focus on the odds of survival. (We can return to odds ratios any time we want simply by forming the ratio of the odds of survival for each of two different levels of SurvRate.)

For what we are doing here (predicting the odds of surviving breast cancer), we will work with the *natural logarithm*[17] of the odds, the result is called the log odds of survival. For our example, the log odds of being delinquent for a male with high testosterone,

$$\log \text{odds} = \log_e(\text{odds}) = \ln(\text{odds}) = \ln(0.293) = -0.228$$

logit

logit transformation

The log odds will be positive for odds greater than 1/1 and negative for odds less than 1/1. (They are undefined for odds = 0.) You will sometimes see log odds referred to as the **logit** and the transformation to log odds referred to as the **logit transformation.**

Returning to the cancer study, we will start with the simple prediction of Outcome on the basis of SurvRate. Letting p = the probability of improvement and $1 - p$ = the probability of nonimprovement, we will solve for an equation of the form:

$$\log(p/1 - p) = \log \text{odds} = b_0 + b_1 \text{ SurvRate}$$

[17] The natural logarithm of X is the logarithm to the base e of X. In other words, it is the power to which e must be raised to produce X, where e is the base of the natural number system = 2.718281.

Here b_1 will be the amount of increase in the *log odds* for a one unit increase in SurvRate. It is important to keep in mind how the data were coded. For the Outcome variable, 1 = no change or worse, and 2 = improvement. For SurvRate, a higher score represents a better prognosis. So you might expect to see that SurvRate would have a positive coefficient, being associated with a better outcome. But with SPSS that will not be the case. SPSS will transform Outcome = 1 and 2 to 0 and 1, and then try to predict a 0 (better). Thus, its coefficient will be negative. (SAS would try to predict a 1, and its coefficient would be positive, though of exactly the same magnitude.)

In simple linear regression, we had formulae for b_0 and b_1 and could solve the equations with pencil and paper. Things are not quite so simple in logistic regression, partly because our data consist of 0 and 1 for SurvRate, rather than the conditional proportions of improvement. For logistic regression, we are going to have to solve for our regression coefficients **iteratively.** This means that our computer program will begin with some starting values for b_0 and b_1, see how well the estimated log odds fit the data, adjust the coefficients, again examine the fit, and so on until no further adjustments in the coefficients will lead to a better fit. This is not something you would attempt by hand.

iteratively

In simple linear regression, you also had standard F and t statistics testing the significance of the relationship and the contribution of each predictor variable. We are going to have something similar in logistic regression, although here we will use χ^2 tests instead of F or t.

In Exhibit 15.6, you will see SPSS results of using SurvRate as our only predictor of Outcome. I am beginning with only one predictor just to keep the example simple. We will shortly move to the multiple predictor case, where nothing will really change except that we have more predictors to discuss. The fundamental issues are the same regardless of the number of predictors.

I will not discuss all the statistics in Exhibit 15.6 because to do so would take us away from the fundamental issues. For more extensive discussion of the various statistics, see Darlington (1990), Hosmer and Lemeshow (1989), and Lunneborg (1994). My purpose here is to explain the basic problem and approach.

The first part of the printout is analogous to the first part of a multiple regression printout, where we have a test on whether the model (all predictors taken together) predicts he dependent variable at greater than chance levels. For multiple regression, we have an F test, whereas here we have (several) χ^2 tests.

Start with the line indicating Beginning Block Number 0, and the row labeled "−2 log Likelihood." At this point there is no predictor in the model and −2 log likelihood = 77.345746. The is a measure of the overall variability in the data. You might think of it as being analogous to SS_{total} in the analysis of variance. The quantity −2 log L can be interpreted as a χ^2 test on how well a model with *no* predictors would fit the data. That χ^2 is 77.3457, which is a significant departure from a good fit, as we would expect with no predictors. (χ^2 would be 0.00 if the fit were perfect.)

For the next block, SPSS adds SurvRate as the (only) predictor and produces another value of −2 log likelihood = 37.323. This is the amount of variability that remains after SurvRate is taken into account, and the difference (77.345 − 37.323 = 40.022) represents a reduction in χ^2 that can be attributed to adding the predictor. Because we have added one predictor, this is itself a χ^2 on 1 *df*, and can be evaluated as such. You can see that the significance level is given as .0000, meaning that SurvRate added significantly to our ability to predict. (You will note that there are lines labeled Model, Block, and Step, and they are all the same because we have added all of our predictors (1) at the same time.)

Number of selected cases: 66
Number rejected because of missing data: 0
Number of cases included in the analysis: 66

Dependent Variable Encoding:

Original Value	Internal Value
1.00	0
2.00	1

Dependent Variable. OUTCOME Cancer Outcome

Beginning Block Number 0. Initial Log Likelihood Function

-2 Log Likelihood 77.345746

* Constant is included in the model.

Beginning Block Number 1. Method: Enter

Variable(s) Entered on Step Number
1.. SURVRATE Survival Rating by Physician

-2 Log Likelihood	37.323
Goodness of Fit	57.235
Cox & Snell - R^2	.455
Nagelkerke - R^2	.659

	Chi-Square	df	Significance
Model	40.022	1	.0000
Block	40.022	1	.0000
Step	40.022	1	.0000

Variables in the Equation

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
SURVRATE	-.0812	.0193	17.7558	1	.0000	-.4513	.9220
Constant	2.6836	.8113	10.9408	1	.0009		

Exhibit 15.6 Logistic analysis of cancer survival

The next section of the table contains, and tests, the individual predictors. (Here, there is only one predictor—SurvRate.) From this section, we can see that the optimal logistic regression equation is

$$\text{Log odds} = -0.0812\ \text{SurvRate} + 2.6836$$

The negative coefficient here for SurvRate indicates that the log odds go down as the physician's rating of survival increases. This reflects the fact that SPSS is trying to predict whether a patient will get worse, or even die, and we would expect that the likelihood of getting worse will decrease as the physician's rating increases.

We can also see that SurvRate is a significant predictor, as tested by Wald's $\chi^2 = 17.7558$ on 1 *df*, which is significant at $p = .0001$ (Wald's χ^2 is a statistic distributed approximately as the chi-square distribution). You will notice that the χ^2 test, that is, $-2 \log L$, on the whole model and the Wald χ^2 test on SurvRate disagree. Because SurvRate *is* the whole model, you might think that they should say the same thing. This is certainly the case

in standard linear regression, where our F on regression is, with one predictor, just the square of our t on the regression coefficient. This disagreement stems from the fact that they are based on different estimates of χ^2. Questions have been raised about the behavior of the Wald criterion, and Hosmer and Lemeshow (1989) suggest relying on the likelihood ratio test ($-2 \log L$) instead.

Looking at the logistic regression equation we see that the coefficient for SurvRate is -0.0812, which can be interpreted to mean that a one point increase in SurvRate will decrease the log odds of getting worse by 0.0812. But you and I probably don't care about things like log odds. We probably want to at least work with odds. But that's easy—we simply exponentiate the coefficient. Don't get excited! "Exponentiate" is just an important sounding word that means "raise e to that power." If you have a calculator that cost you more than $9.99, it probably has a button labeled e^x. Just enter -0.0812, press that button, and you'll have 0.9220. This means that if you increase SurvRate by one point you *multiply* the odds of deterioration by 0.9220. A simple example will show what this means.

Suppose we take someone with a SurvRate score of 40. That person will have a log odds of

$$\text{Log odds} = -0.0812(40) + 2.6837 = -0.5643$$

If we calculate $e^{-0.5643}$ we will get 0.569. This means that the person's odds of deteriorating are 0.569, which means that she is 0.569 times as likely to be deteriorate than improve.[18] Now suppose we take someone with SurvRate = 41, one point higher. That person would have predicted log odds of

$$\text{Log odds} = -0.0812(41) + 2.6837 = -0.6455$$

And $e^{-0.6455} = .524$. So this person's log odds are $-0.6455 - (-0.5643) = -.0812$ lower than the first person's, and her odds are $e^{-0.0812} = 0.9220$ times larger ($0.569 \times 0.922 = .524$). Now, 0.922 may not look like a very large number, but if you had cancer a one point higher survival rating gives you about a 7.8% lower chance of deterioration, and that's certainly not something to sneer at.

I told you that if you wanted to see the effect of SurvRate expressed in terms of odds rather than log odds you needed to take out your calculator and exponentiate. That isn't strictly true here, because SPSS does it for you. The last column in this section is labeled "Exp (B)" and contains the exponentiated value of b ($e^{-0.0812} = .9220$).

Although SurvRate is a meaningful and significant predictor of survivability of cancer, it does not explain everything. Epping-Jordan, Compas, and Howell (1994) were interested in determining whether certain behavioral variables also contribute to how a person copes with cancer. They were interested in whether people who experience a high rate of intrusive thoughts (Intrusiv) have a poorer prognosis. (People who experience intrusive thoughts are people who keep finding themselves thinking about their cancer and related events. They can't seem to put it out of their minds.) These authors were also interested in the effect of avoidant behavior (Avoid), which is exhibited by people who just don't want to think about cancer and who try to avoid dealing with the problem. [Intrusiv and Avoid are variables computed from the Impact of Events Scale (Horowitz, Wilner, and Alvarez, 1979).]

Exhibit 15.7 presents the results of using SurvRate, Intrusiv, and Avoid as predictors of Outcome. You can again see that the overall model fits at better-than-chance levels. With no predictors, -2 log likelihood = 77.346. Adding the three predictors to the model

[18] If you don't like odds, you can even turn this into a probability. Becauses odds $= p/(1 - p)$, then $p = \text{odds}/(1 + \text{odds})$.

Dependent Variable Encoding:

Original Value	Internal Value
1.00	0
2.00	1

Dependent Variable.. OUTCOME Cancer Outcome

Beginning Block Number 0. Initial Log Likelihood Function

-2 Log Likelihood 77.345746

* Constant is included in the model.

Beginning Block Number 1. Method: Enter

Variable(s) Entered on Step Number
1.. SURVRATE Survival Rating by Physician
INTRUS
AVOID

-2 Log Likelihood	31.650
Goodness of Fit	35.350
Cox & Snell - R^2	.500
Nagelkerke - R^2	.724

	Chi-Square	df	Significance
Model	45.695	3	.0000
Block	45.695	3	.0000
Step	45.695	3	.0000

Variables in the Equation

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
SURVRATE	-.0817	.0211	14.9502	1	.0001	-.4092	.9215
INTRUS	-.0589	.0811	.5281	1	.4674	.0000	.9428
AVOID	.1618	.0777	4.3310	1	.0374	.1736	1.1756
Constant	1.6109	1.1780	1.8700	1	.1715		

Exhibit 15.7 Outcome as a function of Survival Rate, Intrusive thoughts, and Avoidance

reduces −2 log likelihood to 31.650, for an improvement of 77.346 − 31.650 = 45.695. This difference is a χ^2 on 3 *df* because we have three predictors, and it is clearly significant. We would have expected a significant model because we knew that SurvRate alone was a significant predictor. From the bottom section of the table, we see that the Wald χ^2 is significant for both SurvRate and for Avoid, but not for Intrusiv. This tell us that people who exhibit a high level of avoidance behavior do not do as well as those who do less avoiding (Wald chi-square = 4.3310, $p = .0374$).[19] More specifically, the regression coefficient for Avoid is 0.1618. This can be interpreted to mean that a one-point increase in Avoid,

[19] In line with Hosmer and Lemeshow's (1989) concern with the validity of the Wald chi-square, we might treat this test with some caution. However, Wald's test tends to be conservative, so confidence in this effect is probably not misplaced. You will see some confirmation of that statement shortly.

holding the other two variables constant, increases the log odds of deterioration by 0.1618 points. Exponentiating this we obtain $e^{0.1618} = 1.1756$. Thus, a one-point increase in Avoid multiplies the odds of detereoration by 1.1756, which would increases them.

The Wald χ^2 test on Intrusiv produced a χ^2 of 0.5281, which was not even close to being significant ($p = .4674$). Thus, this variable is not contributing to our prediction. If Intrusiv is not making a significant contribution of predicting Outcome, perhaps it should be dropped from the model. There is actually a very good reason to do just that. Recall that when we had only one predictor, our overall χ^2, as given by $-2 \log L$, was 40.022. We have now added two more predictors, and our overall χ^2 has become 45.695. The nice thing about χ^2 is that a difference between two chi-squares is itself distributed as χ^2 on *df* equal to the difference between the *df* for the two models. This means that we can compare the fit of the two models by subtracting 45.695 − 40.022 = 5.673 and testing this as a χ^2 on 3 − 1 = 2 *df*. But the critical value of $\chi^2_{.05}(2) = 5.99$, which means that the degree of improvement between the two models is not significant. It is no greater than we would expect if we just added a couple of useless predictors. But we know that Avoid was significant, as well as SurvRate, so what went wrong?

Well, what went wrong is that we have taken the improvement that we gained by adding Avoid, and spread it out over the nonimprovement that we gained by adding Intrusiv, and their average is not enough to be considered significant. In other words, we have diluted the added contribution of Avoid with Intrusiv. If our goal had been to predict Outcome, rather than to test a model that includes Intrusiv, we would have been much better off if we had just stayed with Avoid. So I would suggest noting that Intrusiv does not contribute significantly and then dropping back to the two-predictor model with SurvRate and Avoid, giving us

$$\text{Log odds} = -0.0823\ \text{SurvRate} + 0.1325\ \text{Avoid} + 1.1961$$

Both of these predictors are significant, as is the degree of improvement over the one-predictor case. The fact that adding Avoid leads to a significant improvement in the model over the one-predictor case is welcome confirmation of the significant Wald chi-square for this effect.

The example that was used here included only continuous predictors because that was the nature of the data set. However, there is nothing to preclude dichotomous predictors, and they are often used. The nice thing about a dichotomous predictor is that a one-unit change in that predictor represents a shift from one category to another. For example, if we used Sex as a predictor and coded Male = 1, Female = 2, then a one-unit increase in Sex would move us from Male to Female. The exponentiated coefficient for Sex would then represent the difference in the odds between males and females. Suppose that Sex had been a predictor in the cancer study and that the coefficient was 0.40. Exponentiating this, we would have 1.49. This would mean that, holding all other variables constant, the odds of a female improving are about 1.5 times greater than the odds of a male improving. You will often see statements in the press of the form "Researchers have concluded that people who exercise regularly have a 44% lower chance of developing heart problems than those who do not." Such statements are often based on the kind of reasoning that we are discussing here.

There is much more to logistic regression than I can cover in this short introduction, but perhaps the biggest stumbling block that people experience is the movement to odds and log odds when we are used to thinking about 0 and 1 or about probabilities. My major purpose in this section was to get you past that barrier (and to supply you with arguments why you should consider logistic regression over linear regression or discriminant analysis when you have a dichotomous dependent variable). Everything else that could be said about logistic regression is mainly about the technicalities, and you can find those in a number of texts, particularly the one by Hosmer and Lemeshow (1989).