Chapter 7
Logistic Regression

Linear regression is a widely applicable modeling tool, but it is not appropriate when the correct model should be nonlinear in the parameters. Such is the case when the study endpoint is a binary variable. The model becomes nonlinear because what is being modeled is the probability that a case experiences the event of interest or that a case is in a particular category of the binary response. As a probability must fall between 0 and 1, the linear regression model cannot accommodate it. In this chapter, we examine this important principle, develop the logistic regression model as an alternative, and consider several examples of this modeling strategy from the research literature.

Logistic Regression Model

Often the study endpoint of interest is a binary outcome, for example, whether or not a man's PSA level exceeds 4.0 or whether the result of a prostate biopsy is positive or negative for cancer. When the endpoint is binary, a linear regression model is no longer optimal. Let's consider why. Recall that a linear regression model for the population of units with, say, two regressors for simplicity, takes the form:

$$\mu_y = \alpha + \beta_1 X + \beta_2 Z.$$

Now, in the case of a binary response, Y takes on only two values, which can be represented as 1 if the unit experiences the event of interest and 0 otherwise. The mean of Y is then a proportion; in particular, the proportion of cases experiencing the event of interest in the population, or the *probability* of experiencing the event of interest. Let's let P represent that probability. Then the linear regression model becomes

$$P = \alpha + \beta_1 X + \beta_2 Z.$$

A. DeMaris and S.H. Selman, *Converting Data into Evidence: A Statistics Primer for the Medical Practitioner*, DOI 10.1007/978-1-4614-7792-1_7,

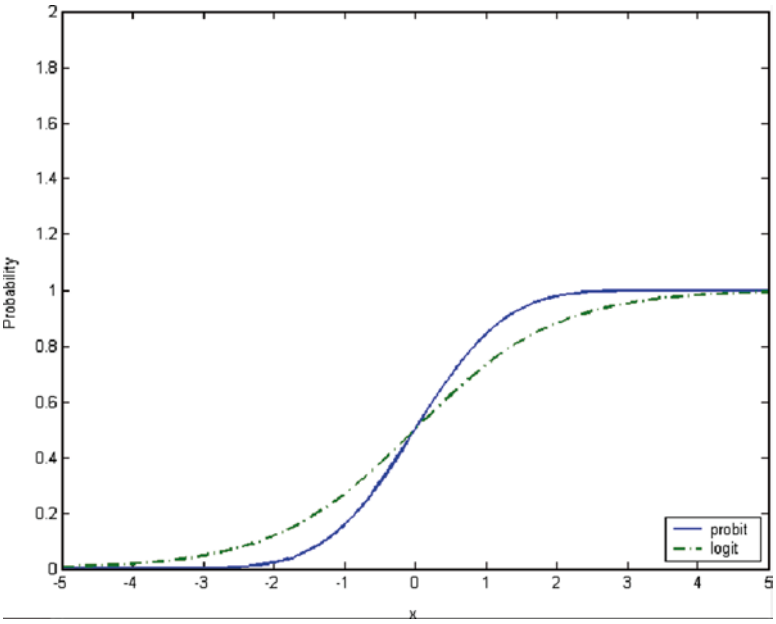

Fig. 7.1 Logit and probit functions giving the probability that a variable "*x*" takes on specific values. Reprinted with permission of John Wiley and Sons, publishers, from DeMaris (2004)

One could estimate such a model with OLS, but it's not the best strategy. The primary problem is that the right-hand side (rhs) of this equation is misspecified. The reason is that a probability has to be within the range 0–1, but the rhs of this equation is not constrained to produce only that range of values. It's entirely possible to get estimated probabilities <0 or >1 with this model. Therefore, a better approach is to find a function for the rhs that is also constrained to stay between 0 and 1. There are two such functions, and they are depicted in Fig. 7.1. The plots show how a probability is related to a single variable, *x*, through these functions.

We see here that, instead of the probability having a linear relationship to *x*—as would be true of the linear regression function—its curve is S shaped, always remaining within the bounds of 0 and 1. The solid line is for the *probit* function, which is used in probit regression, and the dashed line is for the logit function, which is used in logistic regression. We focus only on the logit function in this primer, as it is the preferred technique in medical research for a binary study endpoint. The probit function is used extensively in other fields, such as economics and other social sciences. However, as we shall see, the logit function lends itself to the interpretation of explanatory variable effects in terms of "odds ratios," which is intuitively appealing. The probit function does not have this property. Substantively, however, both modeling techniques result in the same conclusions about the sign and significance of explanatory variable effects on the study endpoint (DeMaris 2004).

The logistic regression model for a probability, as a function of two regressors, is

$$P = \frac{\exp(\alpha + \beta_1 X + \beta_2 Z)}{1 + \exp(\alpha + \beta_1 X + \beta_2 Z)}. \tag{7.1}$$

The rhs here is the algebraic formula that produces the dotted curve in Fig. 7.1 (except the curve in Fig. 7.1 only uses x, rather than $\alpha + \beta_1 X + \beta_2 Z$). In the event that "exp" is not familiar: "exp()" refers to the *exponential function*. "Exp(a)" means to raise Euler's constant to the value of a. Euler's constant is approximately equal to 2.72. For example, exp(2) is $2.72^2 = 7.398$. Euler's constant has a considerable amount of importance in both calculus (Anton 1984) and statistics (DeMaris 2004). The natural logarithm is the inverse function for the exponential function. The natural logarithm of a is the number we have to raise Euler's constant to in order to get a. For example, $\ln(7.398) = 2$ because $2.72^2 = 7.398$. Moreover, $\exp(\ln(a)) = a$ and $\ln(\exp(a)) = a$. Hence, the natural logarithm and exponential functions go hand in hand.

Because the rhs of (7.1) is a complex nonlinear function, it's not very easy to interpret the βs to describe how the regressors affect the probability. But the model can be transformed into a more interpretable version by applying the *logit transformation* to both sides of the equation. The logit transformation of the left-hand side is: $\ln[P/(1-P)]$, where "ln" refers to the natural logarithm (I use "log" interchangeably with "ln" in this primer). Substituting the rhs of (7.1) in place of P in the logit transformation gives us the transformation for the rhs of the equation in (7.1). The result is the logistic regression equation:

$$\log\left(\frac{P}{1-P}\right) = \alpha + \beta_1 X + \beta_2 Z. \tag{7.2}$$

The expression inside the parentheses, $P/(1-P)$, is the *odds* of event occurrence. The odds is just the ratio of two probabilities. (Recall that, although it seems grammatically incorrect, the odds is treated as singular.) In this case it's the ratio of the probability the event occurs to the probability it does not occur. The odds is intuitive for most people. For example, 2-to-1 odds, or an odds of 2, indicates that the event is twice as likely to happen as not. This means that the probability of event occurrence must be 0.667, since $0.667/0.333 = 2$. The left-hand side of (7.2) is therefore the log of the odds (or *log-odds*) of event occurrence. The rhs is the same as in linear regression. So the βs are interpretable as the change in the log-odds per unit increase in a given regressor, holding the other regressor constant. This still isn't entirely satisfactory, since it's hard to get a feeling for what a change in the log-odds means. Therefore, we can write the equation yet once more, in terms of the odds itself:

$$\frac{P}{1-P} = \exp(\alpha + \beta_1 X + \beta_2 Z). \tag{7.3}$$

Now, let's consider what happens to the odds if we increase X by one unit while holding Z constant:

$$\frac{P}{1-P}\Big|_{x+1} = \exp(\alpha + \beta_1(X+1) + \beta_2 Z) = \exp(\alpha + \beta_1 X + \beta_2 Z)\exp(\beta_1). \quad (7.4)$$

What happens is the original odds (which equals $\exp(\alpha+\beta_1 X+\beta_2 Z)$) gets multiplied by $\exp(\beta_1)$. Therefore, we can say that each unit increase in X, holding Z constant, magnifies the odds by $\exp(\beta_1)$. This provides a convenient way to describe how each independent variable affects the odds of event occurrence. We just exponentiate the relevant regression coefficient to find the magnitude of the (multiplicative) change in the odds for a unit increase in the regressor. $\text{Exp}(\beta_1)$ is called the *odds ratio* (since it's the ratio of the odds in (7.4) to the odds in (7.3), above) and is often the preferred way of presenting logistic regression results.

Estimation of Logistic Regression Coefficients

Logistic regression models are not estimated with OLS. Instead, we use one of the most important estimation techniques in statistics: *maximum likelihood estimation*. The way this works is as follows. For any unit sampled from the population, we can express the probability that $Y=1$ for that unit as

$$P(y=1) = P^y(1-P)^{1-y}.$$

This is called the *Bernoulli* probability distribution function. So, for that unit, the probability that his or her y is 1 is $P^1(1-P)^{1-1} = P$. And the probability that his or her y is 0 is $P^0(1-P)^{1-0} = 1-P$. The formula just expresses those probabilities in a compact form. Note that P varies over cases, since people's risks for an event vary from person to person. Now, the probability that we get a particular collection of ones and zeros for our Ys in any sample is just the product of all those Bernoulli functions over all of the sample cases (this is the same principle we use to figure the probability of getting three heads in a row in three coin tosses: it's (0.5)(0.5)(0.5)=0.125). That joint probability is

$$P(\mathbf{y}) = \prod P^y(1-P)^{1-y}. \quad (7.5)$$

where $P(\mathbf{y})$ here represents the probability associated with the complete collection of ones and zeros in the sample, and the large "π" indicates the multiplication together of several terms. The function in (7.5) is called the *likelihood function* for the logistic regression model. Recall that P is a function of the regressors and their effects (i.e., the βs), as shown in (7.1) above. So, substituting (7.1) for P in (7.5) makes it clear that (7.5) is a complex function of the βs. In fact, once the sample has been gathered, the Xs and Ys are fixed. So $P(\mathbf{y})$ in (7.5) is then only a function of α

and the βs. In maximum likelihood estimation, we choose as values for the α and the βs those values that maximize the likelihood function. These are then the parameter values that would have made the researcher's sample of *Y*s *most likely to have been observed* (hence the name "maximum likelihood"). The estimation process is an iterative scheme in which a series of successive approximations is used to find the solution to a collection of nonlinear simultaneous equations. When this solution is found, the parameter estimates can then be plugged back into (7.5) to arrive at an estimated likelihood or probability of observing the sample *Y*s. This is sometimes reported in logistic regression results in log form, i.e., one may see "log likelihood" reported in a logistic regression table. However, this quantity is of no particular interest in and of itself and can be safely ignored.

An Example

Recall the 2002 GSS data used to illustrate multiple linear regression in the previous chapters. In that same survey, respondents were asked about whether or not they had health insurance. Figure 7.2 shows how the question was presented and coded, along with the responses for 2,755 respondents giving valid responses to the question.

Here, "IAP" means inapplicable. We see that the vast majority of respondents—86.6 %—have health insurance. Only 13.4 % of respondents do not have health insurance. (Ten respondents said either that they did not know whether they had health insurance or gave no answer to the question.) What kinds of people don't have health insurance, then? To find this out, we can perform a logistic regression using this variable as the study endpoint. However, we will recode it so that 1 = no health insurance and 0 = has health insurance. The new variable is called "uninsured." Also, we only use the 1,773 respondents who had valid data for all variables in the analysis. The results are shown in Table 7.1.

HLTHPLAN **R HAD MEDICARE OR MEDICAID**

Description of the Variable

855. Do you have any health insurance, including Medicare or Medicaid?

Percent	N	Value	Label
86.6	2,387	**1**	YES
13.4	368	**2**	NO
	52,322	**0**	IAP
	7	**8**	DONT KNOW
	3	**9**	NO ANSWER
100.0	**55,087**		**Total**

Fig. 7.2 Distribution of insured status for GSS respondents

Table 7.1 Logistic regression of uninsured status on explanatory variables in the GSS

Predictor	b	SE(b)	Exp(b)	z	p value	95 % CI
Intercept	2.3205	0.4394	–	5.2811	<0.0001	–
Age	−0.0278	0.0057	0.9730	−4.8772	<0.0001	(0.9620–0.9830)
Female	−0.5581	0.1521	0.5720	−3.6693	0.0002	(0.4250–0.7710)
Education	−0.1265	0.0277	0.8810	−4.5668	<0.0001	(0.8350–0.9300)
Income	−0.0967	0.0132	0.9080	−7.3258	<0.0001	(0.8850–0.9320)
Black	0.0250	0.1980	1.0250	0.1263	0.8994	(0.6960–1.5120)
Other race	0.7558	0.2383	2.1290	3.1716	0.0015	(1.3350–3.3970)
Model χ^2	161.2327				<0.0001	
Df	6					
H-L χ^2	12.1883				0.1430	
Df	8					
Pseudo R_1^2	0.1041					
Pseudo R_2^2	0.2004					

The table shows the explanatory variables used in the model ("Predictor"), the regression coefficients ("b"), the standard errors of the regression coefficients ("SE(b)"), the exponentiated regression coefficients ("Exp(b)"), the test statistic for testing whether each regression coefficient is significant ("z"), the p value for the test statistic ("p value"), and a 95 % confidence interval for the exponentiated regression coefficients ("95 % CI"). Four decimal places are used throughout so that some of the example computations can be illustrated. In the bottom half of the table, beginning with "Model χ^2," are several measures of the goodness of the model that will be explained below.

Interpreting the Coefficients

Several of the individual coefficients are significant. Thus we see that older respondents, women, the more educated, and those with more income are all less likely to fall into the uninsured category. Compared to Whites, however, those of other races than Black (the "Other race" variable) are more likely to be uninsured. On the other hand, Blacks are no different from Whites in the probability of being uninsured, controlling for other factors in the model. The "Exp(b)" column converts the coefficients into odds ratios for easy interpretation. Thus, each additional year of education magnifies the odds of being uninsured by a factor of 0.881. To express this in terms of a percent change in the odds, we use the transformation $100 \times [\exp(b) - 1]$. That is, each year of education reduces the odds of being uninsured by about $100 \times [0.881 - 1] = -11.9$, or 11.9 %. The odds of being uninsured for those of other races is 2.129 times greater than the odds for Whites. Or, the odds of being uninsured for those of other races is $100 \times [2.129 - 1] = 112.9$ % greater than for Whites. The other odds ratios are similarly interpreted.

Predicted Probabilities

Suppose we would like to get the estimated probability of being uninsured, based on the model, for a particular profile of person: a 25-year-old male of race other than Black or White, with a high-school education and average income (mean income is 13.773 here). Here's how we go about it. First, let's get the estimated log-odds of being uninsured for this person by evaluating the equation using their characteristics:

$$\text{Log}\,(P/(1-P)) = 2.32 - 0.028(25) - 0.558(0) - 0.127(12) - 0.097(13.773) + 0.756(1) = -0.484.$$

Then this person's estimated odds of being uninsured is obtained by exponentiating this result:

$$\text{Exp}(-0.484) = 0.616.$$

Finally, the estimated probability of being uninsured is just the odds divided by one plus the odds:

$$P = 0.616/(1+0.616) = 0.381.$$

Hence, according to the model, this person has about a 38 % chance of being uninsured.

Test Statistics and Confidence Intervals

Maximum likelihood estimation assumes that one's sample size is reasonably large. Under that condition, the regression coefficients have a normal distribution. Therefore the test statistic for testing whether each regression coefficient is significant is a z test, just like the test statistic for testing that the population mean is a particular value from Chap. 3. That is, for any regression coefficient, b, the hypothesis is that the corresponding population regression coefficient, β, is zero. Since b is normally distributed we need to find out how many standard deviations b is away from zero so we can know how discrepant the sample results are from what we would expect under the null. The "standard deviation" in question is the standard error of the coefficient, i.e., SE(b). So the test statistic is a z test of the form:

$$z = \frac{b-0}{\text{SE}(b)} = \frac{b}{\text{SE}(b)}.$$

That is, the test statistic is just the ratio of the coefficient to its standard error. For example, from Table 7.1, the test for whether the effect of age is significant is $z = -0.0278/0.0057 = -4.8772$, with a p value that is <0.0001. It is, indeed, very significant. The other z test statistics for the other coefficients are calculated in the same manner.

The confidence intervals in the column "95 % CI" are arrived at using the standard errors of the coefficients, along with the knowledge that the coefficients are normally distributed. Because each coefficient is normally distributed, adding and subtracting 1.96 standard errors from it gives us a 95 % confidence interval for the coefficient. For example, a 95 % confidence interval for the coefficient for being female is $-0.5581 \pm 1.96(0.1521) = (-0.8562, -0.2600)$. This means we are 95 % confident that the effect of being female (vs. being male) on the log odds of being uninsured is between −0.8562 and −0.2600. This can easily be converted into a confidence interval for the odds ratio [Exp(b)] by exponentiating both values. Thus $\exp(-0.8562) = 0.4250$, and $\exp(-0.2600) = 0.7710$, which agrees with the confidence interval shown in the table.

There is also a global test for the utility of the model in logistic regression. This is comparable to the overall F test in linear regression discussed in the previous chapter. If this is significant, then at least one of the coefficients of the regression in the population is nonzero, and we then use the z tests discussed above to discern which these are. The global test for logistic regression, however, is not an F test. Rather it is a chi-squared test and is called the *Model Chi-Squared Test* (or the *Likelihood-Ratio Chi-Squared Test*) and is denoted "Model χ^2" in Table 7.1. As is evident, the test is very significant ($p < 0.0001$), suggesting that the model is of some utility in predicting uninsured status.

Examining Model Performance

Although a model may be of some utility in predicting the study endpoint, we may want to know, in particular, *how much* utility. There are various ways of assessing the model's "fit" to the data or the model's "predictive utility." DeMaris (2004) has labeled model fit *empirical consistency*. This refers to the extent to which the study endpoint "behaves" the way the model says it should. On the other hand, he labels predictive utility *discriminatory power*. This property refers to the extent to which the model is able to separate, or discriminate, different cases' statuses on the study endpoint from each other. Here we discuss measures of both empirical consistency and discriminatory power for the logistic regression model.

Hosmer–Lemeshow Chi-Squared Test. Define a "case" as a subject experiencing the event of interest and a "control" as a subject who does not experience the event. A widely used test of empirical consistency for the logistic regression model is the Hosmer–Lemeshow test (Hosmer and Lemeshow 2000). The idea behind this measure is to use the chi-squared statistic to compare the observed frequencies of cases and controls in the sample with their expected values under the model. With quantitative variables in a logistic regression model, however, each subject typically has a unique predicted probability of being a case. This means that there are as many different predicted probabilities of being a case as there are subjects in the sample. It might seem reasonable to compare whether subjects really are cases with these

Table 7.2 Deciles of risk and Hosmer–Lemeshow chi-squared test of empirical consistency

Partition for the Hosmer and Lemeshow Test					
		Uninsurd = 1		Uninsurd = 0	
Group	Total	Observed	Expected	Observed	Expected
1	177	7	4.78	170	172.22
2	177	4	7.79	173	169.21
3	177	7	10.62	170	166.38
4	177	10	13.12	167	163.88
5	177	16	16.08	161	160.92
6	177	21	19.75	156	157.25
7	177	30	24.48	147	152.52
8	177	36	31.38	141	145.62
9	177	50	42.09	127	134.91
10	180	60	70.92	120	109.08
Hosmer and Lemeshow Goodness-of-Fit test					
Chi-square		DF		Pr > ChiSq	
12.1883		8		0.1430	

probabilities, however, this cannot be done using a chi-squared test. In order to maintain the properties necessary for the statistic to have a chi-squared distribution, subjects are grouped into categories based on their predicted probabilities of being a case. In particular, *deciles of risk* are formed based on the predicted probabilities of being a case. Group 1 consists of the $n/10$ subjects with the lowest probabilities, group 2 the $n/10$ subjects with the next-lowest probabilities, and so on, up to group 10, which consists of the 10 % of the sample with the highest predicted probabilities. Let $\hat{P}$ equal the predicted probability of being a case, according to the model. Once the 10 groups have been identified, the expected number of cases in each group is calculated as the sum of $\hat{P}$ over all subjects in that group. Similarly, the expected number of controls is the sum of $(1-\hat{P})$ over all subjects in the same group. The Hosmer–Lemeshow statistic is then the chi-squared statistic for the resulting table of observed and expected frequencies. Under the null hypothesis that the model is empirically consistent, this statistic has a chi-squared distribution with 8 degrees of freedom. A significant χ^2 implies a model that is *not* empirically consistent. Table 7.2 shows the deciles of risk and the ensuing Hosmer–Lemeshow chi-squared test for empirical consistency for the logistic regression model in Table 7.1. This table was produced by the SAS software program.

The table shows the deciles as the "Group" column. The first decile, consisting of 177 subjects, is the group with the lowest risks of being uninsured, according to the model. It has 7 observed cases and 170 observed controls. According to the model, the expected number of cases for this group is 4.78, and the expected number of controls is 172.22. The other deciles all have higher risks of being uninsured, culminating in Group 10, the highest decile of risk, with 180 subjects. For this group, there were 60 observed cases and 120 observed controls. The expected number of cases and controls in this group, according to the model, are 70.92 and 109.08,

respectively. The Hosmer–Lemeshow statistic is shown at the bottom of the table as 12.1883. It is not significant ($p = 0.1430$). This means that the expected numbers of cases and controls (according to model predictions) are not very different from the actual numbers of cases and controls. And this suggests that the model is indeed empirically consistent or has an acceptable fit to the data. This statistic is also reported in the bottom half of Table 7.1 as "H-L χ^2." However, an empirically consistent model may not have much predictive power, as the following discussion reveals.

Pseudo-R^2 Values. In multiple linear regression, the most commonly used measure of discriminatory power is R^2. In logistic regression, because of the binary nature of the study endpoint, calculating an R^2 measure is far more complicated. Many counterparts to R^2 have been proposed for use in logistic regression (see, for example, Long 1997), but no single measure is consistently used. Additionally, many of these do not have the same interpretation as in linear regression. Although they typically range from 0 to 1, they cannot be interpreted as the variance in the study endpoint explained by the model. In an extensive simulation, DeMaris (2002) investigated the performance of eight popular pseudo-R^2 measures for logistic regression. The two best-performing measures are shown in Table 7.1 as "Pseudo R_1^2" and "Pseudo R_2^2." An advantage to these two measures is that both of them do have an explained-variance interpretation. However, they differ as to what the study endpoint represents. Pseudo R_1^2 (referred to as "explained risk" and denoted "$\hat{\Delta}$" by DeMaris) assumes that the study endpoint is a true qualitative difference in state. In this example, that's reasonable. Either one has health insurance or one does not. A woman is either pregnant or she is not. And so forth. Pseudo R_1^2 is then interpreted as the variation in the event in question that is accounted for by the logistic regression model. In the current example, it's telling us that about 10 % of whether or not one has health insurance is explained by the model. Pseudo R_2^2, on the other hand (called the "McKelvey-Zavoina R^{2}" and denoted "R_{MZ}^2" by DeMaris), is more appropriate when the binary study endpoint is a crude proxy for a quantitative underlying variable. For example, suppose we are studying depressive symptomatology. Subjects have all taken the CES-D and have a score on depressive symptomatology as a result. But the only information retained for them is whether or not their score was >25, a threshold deemed the cutoff for being clinically depressed. So all we have recorded on subjects is a binary indicator of whether or not they are clinically depressed. In a logistic regression of this binary indicator on a set of predictors, our interest might be in how the predictors influence depressive symtomatology per se, not just whether someone is clinically depressed. In that case, we might want to estimate the variance explained by our model in the quantitative underlying variable of depressive symptomatology. Pseudo R_2^2 would be the measure to use for this. Thus, if whether or not one has health insurance were a proxy for a quantitative measure of the *extent* of health insurance, say, then Pseudo R_2^2 is telling us the model explains about 20 % of the variance in that underlying measure. As a final note, neither Pseudo R_1^2 nor Pseudo R_2^2 is a routine part of the

Table 7.3 Classification table for being uninsured, based on logistic regression model in Table 7.1

	Observed status		
Classified	Insured	Uninsured	Total
Insured	1,519	223	1,742
	99.2 %	**92.5 %**	
Uninsured	13	18	31
	0.8 %	**7.5 %**	
Total	1,532	241	1,773
	86.4 %	**13.6 %**	
Criterion	0.50		
Sensitivity	7.5 %		
Specificity	99.2 %		
False positive rate	0.8 %		
Percent correctly classified	86.7 %		
Percent correct by chance	76.5 %		

output of statistical software. For this reason, they are not yet commonly used. So if the reader sees a "Pseudo R^{2}" measure reported for logistic regression, he or she should not assume that it has an explained-variance interpretation.

The ROC Curve. Another way to examine discriminatory power for the logistic regression model is to examine how well it allows us to correctly classify subjects with respect to the study endpoint. This is assessed in the following manner. Obtain the model-predicted probabilities of experiencing the event for each subject, in the manner illustrated above for our 25-year-old male. If that probability is greater than some criterion value, typically taken to be 0.50, classify that subject as a case. If the probability is below the criterion, classify that subject as a control. Then compare the model-based classification to the subject's actual status on Y to see how well the model leads to correct prediction of the subject's status on Y. Repeat this operation for all the subjects in the sample. Table 7.3 shows the result of this process for the logistic regression model in Table 7.1.

We see that, of 241 uninsured cases in the sample, 18 or 7.5 % were correctly classified as uninsured by the model. The probability of a case being classified by the model as a case is called the *sensitivity* of classification; therefore sensitivity is 7.5 % for this model. On the other hand, the probability of a control being classified by the model as a control is called the *specificity* of classification. In this example, 1,519 out of 1,532 controls were correctly classified as controls. Therefore, specificity is 99. 2 %. One minus the specificity is the *false positive rate*, i.e., the probability of a control being mistakenly classified by the model as a case. In this instance, that is 0.8 %. To the extent that sensitivity is greater than the false positive rate, as in this instance, the model has value. The probability of a case being classified as a case is greater than the probability of a control being classified as a case. On the whole, however, the model doesn't appear to perform all that well, which is also consistent with the relatively low pseudo-R^{2} values in Table 7.1. In all, 1,519 insured subjects

Table 7.4 Classification table for being uninsured, based on logistic regression model in Table 7.1

	Observed status		
Classified	Insured	Uninsured	Total
Insured	1,431	190	1,621
	93.4 %	**78.8 %**	
Uninsured	101	51	152
	7.6 %	**21.2 %**	
Total	1,532	241	1,773
	86.4 %	**13.6 %**	
Criterion	0.30		
Sensitivity	21.2 %		
Specificity	93.4 %		
False positive rate	6.6 %		
Percent correctly classified	83.6 %		
Percent correct by chance	76.5 %		

were correctly classified as "insured" by the model, and 18 uninsured subjects were correctly classified as "uninsured." That means that $(1{,}519+18)/1{,}773=0.867$ or 86.7 % of the cases are correctly classified by the model. However, fully 76.5 % would be correctly classified just by chance alone. But most of the errors in classification are for cases. Perhaps classification performance of the model can be improved by setting the criterion lower.

Table 7.4 shows the results of setting the criterion at 0.30 instead of 0.50.

What this table shows is that sensitivity has been improved, but at the cost of specificity. Sensitivity is 21.2 % but specificity has dropped to 93.4 %. Nevertheless, sensitivity is higher than the false positive rate of 6.6 % But the percent correctly classified has also dropped some to 83.6 %. We notice, however, that we are not misclassifying the cases as badly as we were in Table 7.3, so there appears to be some improvement in that regard.

Since the sample percent uninsured is only 13.59 %, why not try that value as the criterion? Table 7.5 shows this result.

Once again, we have improved sensitivity at the expense of specificity, with both values now approximately the same—69.7 % and 69.5 %, respectively. And we note that sensitivity is more than twice as great as the false positive rate, as well. However, this time only 69.5 % of cases are correctly classified, which is actually worse than we could do by chance alone! Nevertheless, the accuracy of classification of both cases and controls appears to be strongly affected by choice of criterion value.

The idea of varying the classification criterion—as in Tables 7.3, 7.4, and 7.5—gives rise to the *receiver operating characteristic*, or ROC, curve. The idea is to vary the criterion incrementally from 0 to 1, each time generating a classification table such as Tables 7.3, 7.4, and 7.5. Afterwards, a plot of sensitivity against the false positive rate, based on the entire collection of classification tables, produces the ROC curve. This is shown in Fig. 7.3 for the model in Table 7.1.

The area under the curve, or AUC, is the key measure of interest. (This is also called the "concordance index" or the "C" statistic.) It is interpreted as the *likelihood*

Table 7.5 Classification table for being uninsured, based on logistic regression model in Table 7.1

	Observed status		
Classified	Insured	Uninsured	Total
Insured	1,065	73	1,138
	69.5 %	**30.3 %**	
Uninsured	467	168	635
	30.5 %	**69.7 %**	
Total	1,532	241	1,773
	86.4 %	**13.6 %**	
Criterion	0.1359		
Sensitivity	69.7 %		
Specificity	69.5 %		
False positive rate	30.5 %		
Percent correctly classified	69.5 %		
Percent correct by chance	76.5 %		

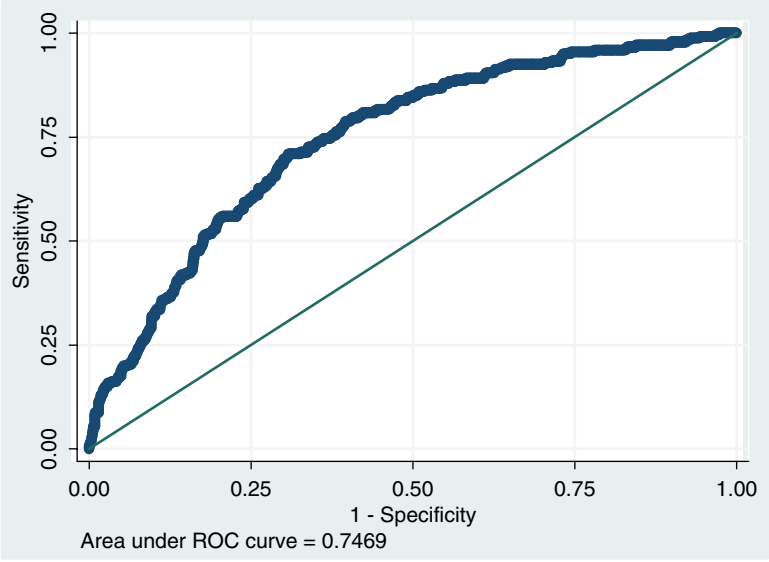

Fig. 7.3 ROC curve for logistic regression model of uninsured status in Table 7.1

that a case will have a higher predicted probability of the event than a control across the range of criterion values investigated. The diagonal line in the middle of the graph represents an AUC of 0.50. This is the minimum AUC a model could demonstrate and would suggest a model of absolutely no discriminatory power. AUC values above 0.7 generally indicate models with acceptable discriminatory power, with higher AUCs implying even better performance. For example, an AUC above 0.80 is considered "excellent," and an AUC above 0.90 is "outstanding"

(Hosmer and Lemeshow 2000). The AUC for the model of uninsured status in Table 7.1 is 0.75, which is just in the adequate range, but not great. This suggests that the model needs improvement before it would be useful for forecasting.

As a final comment, it should be noted that logistic regression is also used when the study endpoint has more than two categories. If these categories represent a qualitative variable, the procedure is then called *multinomial logistic regression*. If the categories represent rank order on some attribute but there are not enough categories to treat the response as a quantitative variable for linear regression, then the technique is called either *ordered logit modeling* or *ordinal logistic regression*. These variants on the logistic regression model see extensive use in the social and behavioral sciences, but are not often employed in medical research.

Applications: Logistic Regression in Action

Logistic regression is an extremely popular tool in medical research. Below we present several examples of interesting applications of the technique to different medical issues.

Morbidity Following Kidney Surgery

Abouassaly et al. (2011) studied the effect of patient age on the morbidity of kidney surgery associated with renal cell carcinoma. They were concerned that previous studies, largely based on single-institution populations, have painted too sanguine a picture about outcomes for this patient population. In their words (p. 812): "Better assessment of surgical morbidity, particularly in those at highest risk, i.e., elderly patients, would allow better preoperative counseling and may suggest the need for less invasive therapy in these groups, e.g., active surveillance or ablative therapy." They employed a database of patients treated between 1998 and 2008, containing information on all acute care renal hospitalizations in nine of the ten Canadian provinces. They excluded pediatric patients, as well as anyone treated for other than a solid or cystic renal mass, leaving a total of 24,578 patients for analysis. Explanatory variables included patient age, Charlson score (a measure of comorbidity), year (coded as fiscal year category), surgeon and hospital volumes for kidney procedures (both coded in quartiles), and patient income level (coded in quintiles). The study endpoint for the logistic regression was the probability of the patient having any complication after surgery. Table 7.6 is a partial reproduction of their logistic-regression results table (results for complications after partial nephrectomy as another study endpoint, as well as some covariates, are not shown).

The results shown here are for the case of radical nephrectomy surgery. Notice that all effects are for qualitative factors, represented as sets of dummy variables. There is just one p value reported in the column "Overall p Value" for each qualitative factor. This p value tells us whether that qualitative factor, per se, has a significant

Table 7.6 Logistic regression analysis of predictors of complications after radical nephrectomy (RN)

	RN*	
	OR (95 % CI)	Overall p Value
Age category		<0.0001
Less than 50	Referent	
50–59	0.98 (0.88–1.08)	
60–59	1.14 (1.03–1.25)	
70–79	1.39 (1.26–1.53)	
80 or greater	1.74 (1.52–1.98)	
Charlson category		<0.0001
0	Referent	
1	1.88 (1.73–2.05)	
2	3.57 (3.19–4.00)	
3 or greater	6.22 (5.18–7.48)	
Fiscal yr category		<0.0001
1998–1999	Referent	
2000–2001	0.99 (0.90–1.09)	
2002–2003	1.04 (0.94–1.15)	
2004–2005	0.95 (0.86–1.05)	
2006–2007	0.68 (0.61–0.75)	
Surgeon vol quartile		<0.0001
Low	Referent	
Intermediate	0.83 (0.77–0.91)	
High	0.76 (0.69–0.83)	
Very high	0.82 (0.74–0.91)	
Hospital vol quartile		<0.0001
Low	Referent	
Intermediate	1.02 (0.94–0.91)	
High	1.41 (1.28–1.55)	
Very high	1.41 (1.28–1.56)	
Income quartile		0.039
Very Low	Referent	
Low	1.06 (0.96–1.17)	
Intermediate	1.15 (1.04–1.27)	
High	1.15 (1.03–1.27)	
Very high	1.13 (1.02–1.26)	

Reprinted with permission of Elsevier Publishers from Abouassaly et al. (2011)

*C-statistic = 0.66, Hosmer–Lemeshow $p = 0.044$

†C-Statistic = 0.65, Hosmer–Lemeshow $p = 0.73$

effect on the risk of complications. If it does, then we would want to know which categories of that factor are "significant." Each category having a coefficient associated with it is being compared to the reference group (labeled "Referent" in the table) for the dummy variables representing that factor. All effects are being reported as odds ratios (OR), with 95 % confidence intervals for the odds ratios in parentheses.

For example, being 80 or older is associated with odds of complications that are 1.74 times higher than for those who are under 50 (the reference group), controlling for the other factors in the model. Or, those 80 or older have 74 % greater odds of

complications compared to those under 50. Those having a Charlson score of 3 or greater have 6.22 times greater odds of developing complications, compared to those with a Charlson score of zero, and so forth. Whether each of these comparisons of a category of a factor with the reference group for that factor is significant can be discerned from the confidence interval for its odds ratio. If that confidence interval does not contain 1.0, then that odds ratio is significant. Returning to our two examples, we see that the confidence interval for the OR for age 80 or greater is 1.52–1.98. This interval does not contain 1.0, so it is significant. This means that the odds of complications for those aged 80 or over are significantly greater than for those aged less than 50. Or, the confidence interval for the OR for a Charlson score of 3 or greater is 5.18–7.48. Again, this interval does not contain 1.0, so this OR is significant. Those with a Charlson score of 3 or greater have significantly greater odds of complications, compared to those with a Charlson score of zero. What do we do if we want to know whether those with a Charlson score of 3 or greater have greater odds of complications than those with Charlson scores of 2 (which is not the reference group)? What the analyst has to do is simply to change the reference group to those with a Charlson score of 2 and rerun the model. Then the OR for those with a Charlson score of 3 or greater will be with reference to those with a Charlson score of 2. This latter comparison may or may not be of interest. We see that several of the ORs are not significant, because their CIs do contain 1.0: the ORs for fiscal years 2000–2005, the OR for the intermediate hospital-volume quartile, and the OR for the low income quintile fall into this category.

Measures of empirical consistency and discriminatory power are reported at the end of the table. The starred (*) entries are for radical nephrectomy (the other two entries are for partial nephrectomy, whose results are not shown). We see that the AUC ("C-statistic") is only 0.66. This is not considered acceptable discriminatory power for a logistic regression model. We notice, too, that the p value for the Hosmer–Lemeshow chi-squared is just significant, at $p = 0.044$. This also suggests a model that does not have a particularly good fit to the data. That very significant explanatory variable effects can coexist with a marginally performing model here is due to the very large sample size. In this case, there is a considerable amount of power for detecting "significant" effects, even though model performance is less than impressive on the whole.

Caffeine, Smoking, and Parkinson Disease

Coffee drinking and cigarette smoking have both been shown, in a number of studies, to be associated with a lower risk of developing Parkinson disease, or PD (Liu et al. 2012). The authors explain the connection of Parkinson's with caffeine (p. 1200): "It has been hypothesized that caffeine and its major metabolites may protect dopaminergic neurons by antagonizing adenosine A2A receptor." With this in mind, Liu and colleagues undertook an evaluation of the influence of caffeine intake and smoking on the development of PD in a large cohort of men and women. They utilized data from the NIH-AARP Diet and Health Study on AARP members

aged 50–71 from six US states and two metropolitan areas. A baseline survey on diet and lifestyle, including coffee and cigarette consumption, was answered in 1995–1996. Then a follow-up survey was conducted in 2004–2006 among surviving participants to ascertain the occurrence of major chronic diseases such as Parkinson's. After excluding cases with missing data, the sample size was 304,980 participants, 1,100 of whom had been diagnosed with PD during or after the year 2000. Caffeine intake was assessed at least 4 years before PD diagnosis for these individuals. In studies without random assignment to levels of the explanatory variables, an important means of control to ensure causal priority is to exclude certain cases. The authors explain (p. 1201): "Because caffeine intake was assessed in 1995–1996 and we were concerned that PD patients might have altered their coffee consumption, even prior to PD diagnosis, we excluded 1,094 potential cases diagnosed before 2000 from the analyses." That is, for individuals diagnosed too close to baseline (1995–1996), at which coffee consumption was measured, developing PD might actually have caused an increase in their coffee consumption. Since coffee consumption is presumed to be a cause of level of risk for PD, these patients demonstrating reverse causation had to be excluded from the study. The statistical analysis consisted of a logistic regression of PD (coded 1 if the respondent had PD, 0 otherwise) on caffeine intake plus control variables.

Participants with higher caffeine intake were more likely to be male, Caucasian, and less physically active. Caffeine intake was strongly associated with cigarette smoking. Higher coffee consumption was associated with a lower risk for PD. But once cigarette smoking was controlled in the analysis, this effect only held for caffeinated coffee. Moreover, consumption of other caffeinated beverages (e.g., tea, soft drinks) was not related to the risk of PD (Liu et al. 2012). The principal findings are explained by the authors (p. 1204) and shown in Fig. 7.4:

> Duration of smoking was strongly associated with lower PD risk; further adjustment for caffeine intake barely changed the risk estimates for smoking (Web Table 3). Joint analysis of smoking duration and caffeine intake showed that smoking was associated with lower PD risk within each level of caffeine intake (Figure 1; for all subgroups, Ptrend ≤ 0.01). In contrast, higher caffeine intake was significantly associated with lower PD risk among never smokers (Ptrend $= 0.04$), but the monotonic trend was less clear among ever smokers. Nevertheless, compared with never smokers with low caffeine intake, long-term smokers with high caffeine intake had the lowest risk of PD. The statistical test for a potential interaction between smoking and caffeine intake was far from statistically significant ($P = 0.57$).

As the authors made clear, there is no interaction between smoking and caffeine intake in their effects on the probability of developing PD. The effects of caffeine intake and smoking appear to be cumulative in reducing the risk for PD, with the lowest odds of developing PD shown by the group with the last bar on the right in the figure. This is the group with high caffeine intake who are either past smokers who smoked for 30 or more years or who are current smokers. Their odds ratio of 0.48 is comparing their odds of PD to those on the far left (with an OR = 1.00), the reference category. Thus, the lowest risk group for PD has odds of developing PD that are only about half (i.e., 0.48) the odds of those with low caffeine intake who never smoked. Controlling for age at baseline, race, physical activity, and gender do not alter these findings.

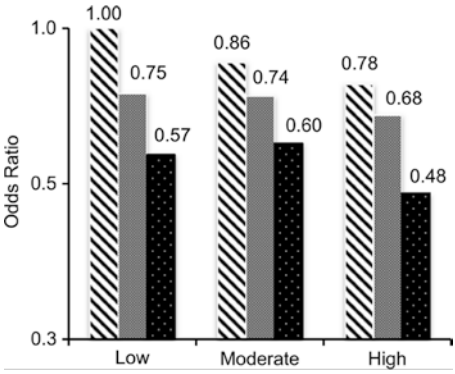

Fig. 7.4 Odds ratios for PD according to caffeine intake (low, moderate, or high) and smoking status (*Striped bars* = Never; *Gray bars* = Past Smoker for 1–29 Years; *Dotted bars* = Past Smoker for ≥30 Years or Current Smoker). Reprinted with permission of Oxford University Press from Liu et al. (2012)

PSA as a Predictor of Prostate Cancer

Crawford and colleagues (2011) conducted a nonexperimental study to determine the prognostic value of initial PSA levels in men for identifying the risk of developing prostate cancer (PC). Their contention is that men with a *first* PSA reading between 1.5 and 4.0 face the same future risk of PC as those with a PSA level above 4.0 in any given examination (Crawford et al. 2011). Their database consisted of men in the Health Alliance Plan of Henry Ford Health System between 1997 and 2008. They were at least 40 years old, had initial PSA values between 0 and 4.0 ng/mL, and had a minimum of 4years of follow-up after their first PSA. As in the previous study, exclusionary criteria were employed to exercise control over direction of causality (p. 1744):

> To assess the future predictive value of a first PSA test, patients could not have been in the system for less than 6 months (to rule out the possibility of referral for prostate cancer) and patients could not have received a diagnosis of prostate cancer within 6 months of baseline PSA (otherwise, possibly representing the PSA that initiated biopsy and diagnosis). These exclusionary criteria were designed to ensure temporal separation between the baseline PSA and a subsequent diagnosis of cancer.

The study endpoint was a diagnosis of PC, coded 1 for such a diagnosis, and 0 otherwise. This was then analyzed via logistic regression using initial PSA value as the primary predictor. Initial PSA value was dichotomized as <1.5 vs. 1.5–4. The authors' description of their analytic technique is instructive (p. 1744):

> Multivariate analysis, adjusting for age and race, was performed using SAS v9.1.3. Initially, the relative risk of prostate cancer was determined for all subjects based on a PSA threshold of 1.5 ng/mL. The PSA threshold analysis was subsequently stratified by race, controlling for age. To determine optimal PSA threshold, receiver operating characteristic curves were constructed and then the sums of sensitivity and specificity were evaluated. Area under the receiver operating characteristic curve (AUC) was used to determine the predictive ability of PSA values for prostate cancer. A perfect test has an AUC of 1.0, whereas a test with no diagnostic value has an AUC < 0.5.

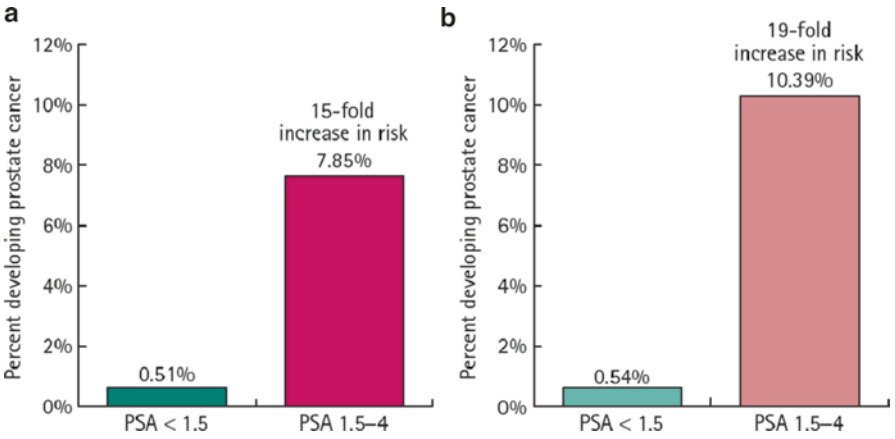

Fig. 7.5 Risks of prostate cancer for the entire sample (**a**) and African-Americans Only (**b**). Reprinted with permission of John Wiley and Sons, publishers, from Crawford et al. (2011)

We notice that the authors mention the statistical software package they used to analyze the data as being SAS v 9.1.3. SAS output has been shown in previous chapters. We see, also, that some analyses were "stratified" by race, that is, analyses were run separately for different racial groups, and included age as a control variable. Apparently the authors explored different PSA cutoffs for the dichotomized PSA explanatory variable but found that 1.5 provided the greatest AUC and the best values of sensitivity and specificity. Moreover, AUC was used to assess discriminatory power of the model, as has been illustrated above for the GSS example.

The primary study findings are illustrated in Fig. 7.5.

As illustrated in the figure, and emphasized by the authors, men with a baseline PSA $\geq$ 1.5 ng/mL had odds of prostate cancer that were 15 times higher than those with PSA < 1.5 ng/mL. For African-American men, those with PSA $\geq$ 1.5 ng/mL had a PC risk that was 19 times higher than those with PSA < 1.5 ng/mL. How good was the researchers' logistic regression model for forecasting PC? The AUC results are depicted in Fig. 7.6.

The figure shows that the AUC was 0.873, which suggests excellent discriminatory power for the authors' model. Sensitivity and specificity, according to the figure, are both about 0.80.

Vitamin D Deficiency and Frailty

Another study employing logistic regression and reporting the AUC for the model is by Wilhelm-Leen et al (2010). Their primary study endpoint was frailty in older persons, described as a "multidimensional phenotype that describes declining physical function and a vulnerability to adverse health outcomes in the setting of

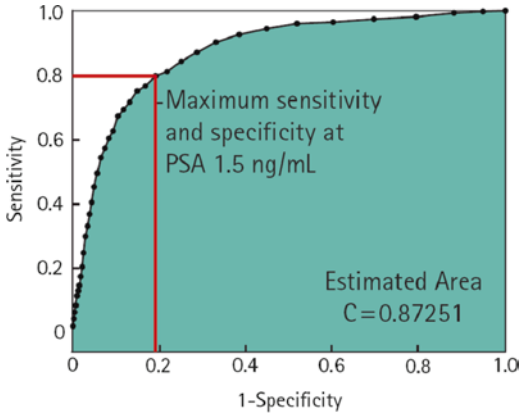

Fig. 7.6 Receiver operating characteristic curve for PC prediction for all study patients. Reprinted with permission of John Wiley and Sons, publishers, from Crawford et al. (2011)

physical stress such as illness or hospitalization" (Wilhelm-Leen et al., p. 171). Their hypothesis was that 25-hydroxyvitamin D deficiency would be predictive of frailty in older adults, controlling for advanced age and chronic medical conditions. They utilized data from the Third National Health and Nutrition Evaluation Survey, a nationally representative survey of the health status of persons residing in the USA collected in the period 1988–1994. Their sample consisted of 5.048 persons aged 60 or older with 25-hydroxyvitamin D data available. Frailty was coded as 1 for frail, 0 for not frail. This was based on respondents having three or more of the following conditions: low body weight for height, slow walking, weakness, exhaustion, and low physical activity. The authors controlled for several factors in their analysis, such as age, sex, and, poverty status, and various comorbidities, such as diabetes, chronic lung disease, and chronic kidney disease. The logistic regression results for Whites are shown in Table 7.7.

In the original table title (not reproduced here), AUC was reported as 0.767. This indicates a model with acceptable discriminatory power. We see, also, that the primary explanatory variable, vitamin D, has the expected effect. Those with D level less than 15 ng mL^{-1} have an estimated odds of being frail that is over three times greater (3.7, to be exact) than those with D levels greater than or equal to 30 ng mL^{-1}. And this holds while controlling for age, gender, poverty-to-income ratio (PIR), and various comorbidities. As for the other factors, there appears to be no significant gender difference in the risk of frailty. But not surprisingly, the probability of being frail increases with age, a lower PIR, and the conditions of arthritis, nonskin cancer, chronic kidney disease, cardiovascular disease, and diabetes.

Heat Sensitivity in MS Patients

Sensitivity to environmental heat is a well-known concomitant of multiple sclerosis (MS) that exacerbates MS symptoms. Flensner et al. (2011) examined the effects of

Table 7.7 Logistic regression results for frailty of white respondents

	OR	95 % CI
Vitamin D (ng mL^{-1})		
≥30	Reference	–
15–<30	1.0	0.6–1.7
<15	3.7	2.1–6.8
Age (years)		
60–69	Reference	–
70–79	1.9	1.3–2.8
≥80	2.5	1.4–4.5
Sex		
Male	Reference	–
Female	1.2	0.8–1.8
Poverty to income ratio (PIR)		
PIR≥2	Reference	–
PIR<2	1.9	1.3–2.6
Comorbidity		
Arthritis	3.8	2.2–6.5
Cancer, nonskin	1.9	1.2–2.9
Chronic liver disease	1.4	0.7–2.7
Chronic lung disease	1.4	0.8–2.3
Chronic kidney disease	1.7	1.1–2.6
Cardiovascular disease	1.8	1.2–2.6
Diabetes	1.6	1.1–2.3

Reprinted with permission of John Wiley and Sons, publishers, from Wilhelm-Leen et al. (2010)

heat sensitivity on a variety of common MS symptoms. Their data were drawn from 334 MS sufferers in the Swedish MS Register. Inclusion criteria were being diagnosed with MS, having an Expanded Disability Status Score (EDSS) between 0 and 6.5, and being between 20 and 65 years of age (Flensner et al. 2011). Information was gathered from respondents via mailed questionnaires. Heat sensitivity was based on a single question: "Are you sensitive to heat?" (Flensner et al. (2011), p. 2). This was coded simply "yes" (1) and "no" (0). Table 7.8 presents logistic regression results for the effects of heat sensitivity and the EDSS score on several MS symptoms. The authors' table title notes that each symptom is coded "1 = never to sometimes, 2 = usually to always." We should note that, although the study endpoint for logistic regression is usually coded 1 and 0, this coding is not a requirement. Any two numerical codes will suffice, provided they are recognized by the software used to analyze the data.

As is evident, heat sensitivity has significant effects on six of the eight symptoms shown. In all cases, heat sensitivity exacerbates the symptom. For example, those who are heat sensitive have odds of fatigue that are about two-and-a-half times greater than those who are not heat sensitive. Similar effects are seen for leg weakness, concentration difficulties, pain, paraesthesia, and urination urgency. EDSS is also associated with several symptoms. Unique to this analysis is the reporting of a

Table 7.8 Logistic regression analysis of common MS symptoms on EDSS score and heat sensitivity

	EDSS			Heat sensitivity			
MS symptoms	OR	95 % CI	*P*-value	OR	95 % CI	*P*-value	R^2 Nagelkerke
Fatigue	1.15	0.98–1.32	0.086	2.55	1.48–4.25	<0.001	0.136
Leg weakness	1.51	1.26–1.81	<0.001	2.21	1.24–3.93	0.007	0.274
Spasms	1.79	1.43–2.22	<0.001	1.65	0.77–350	0.194	0.232
Balance problems	1.62	1.34–1.94	<0.001	1.48	0.83–2.65	0.181	0.285
Concentration difficulties	1.08	0.92–1.28	0.354	3.40	1.85–6.25	<0.001	0.123
Pain	1.09	0.92–1.29	0.344	3.55	1.87–6.77	<0.001	0.136
Paraesthesia	1.20	1.02–1.41	0.026	2.10	1.21–3.64	0.008	0.095
Urination urgency	1.27	1.05–1.54	0.016	2.75	1.28–5.90	0.009	0.256

Reprinted from Flensner et al. (2011), an open-access journal

pseudo-R^2 value: "R^2 Nagelkerke." For each MS symptom, R^2 pertains to the logistic regression model containing two predictors: EDSS and heat sensitivity. The Nagelkerke R^2 is similar to Pseudo R_2^2 in Table 7.1 and discussed above. It is a good estimate of the quantitative variable that underlies a binary indicator. In this case, in which the study endpoints refer to the frequency or intensity of MS symptoms, such a quantitative underlying variable is quite plausible. The advantage to the Nagelkerke R^2 is that it is frequently reported as a standard part of logistic regression software. The disadvantage is that, unlike the linear regression R^2, Pseudo R_1^2, and Pseudo R_2^2, it does not have an explained-variance interpretation (DeMaris 2002). It simply indicates the degree of discriminatory power of the model, on a scale from 0 to 1. Apparently, the model demonstrates the greatest predictive efficacy for the study endpoint "balance problems."

This chapter has dealt primarily with binary logistic regression, a technique that is appropriate whenever we have a dichotomous outcome variable. But what should we do if we have a dichotomous outcome but it represents an event that occurs to cases that are followed longitudinally? For example, we might follow patients from the time of their diagnosis with a potentially fatal disease to see what factors affect whether they die. It turns out that we do not just want to perform a logistic regression with death as the binary outcome as our analytic strategy. The reason is that we want to take account of *how long they survive until death*, not just whether they die or not. There will also be patients who are still alive at the end of the observation period. These patients have survival times that are said to be "censored." Rather than just treat these cases as though they are "safe," we incorporate the censoring into the analyses. These nuances of time-to-event data are all readily incorporated into the technique called *survival analysis*, the subject of the next chapter.