Chapter 15

Logistic Regression: Modeling Categorical Responses

The regression models studied in the past six chapters assume that the response variable is quantitative. This chapter presents generalized linear models for response variables that are categorical.

Sections 15.1 – 15.3 present the ***logistic regression model*** for ***binary*** response variables — variables having only two possible outcomes. For instance, logistic regression can model

- A voter's choice in a presidential election (Democrat or Republican), with predictor variables political ideology, annual income, education level, and religious affiliation.
- Whether a person uses illegal drugs (yes or no), with predictors education level, whether employed, religiosity, marital status, and annual income.

Multi-category versions of logistic regression can handle ordinal response variables (Section 15.4) and nominal response variables (Section 15.5). Section 15.6 introduces ***loglinear models***, which describe association structure among a set of categorical response variables. Section 15.7 shows how to check the goodness of fit of models to contingency table data. The models of this chapter use the ***odds ratio*** to summarize associations.

15.1 Logistic Regression

For a binary response variable y, denote its two categories by 1 and 0. Commonly the generic terms *success* and *failure* are used for these two outcomes. Recall (from the discussion of Table 3.6 on page 61 and Example 4.8 on page 120) that the mean of 0 and 1 outcomes equals the proportion of outcomes that equal 1. Regression models for binary response variables describe the population proportions. The population proportion of successes also represents the probability $P(y = 1)$ for a randomly selected subject. This probability varies according to the values of the explanatory variables.

Models for binary data ordinarily assume a *binomial distribution* for the response variable (Section 6.7). This is natural for binary outcomes. The models are special cases of generalized linear models (Section 14.4).

Linear Probability Model

For a single explanatory variable, the simple model

$$P(y = 1) = \alpha + \beta x$$

implies that the probability of success is a linear function of x. This is called the ***linear probability model***.

This model is simple but often inappropriate. As Figure 15.1 shows, it implies that probabilities fall below 0 or above 1 for sufficiently small or large x values, whereas probabilities must fall between 0 and 1. The model may be valid over a restricted range of x values, but it is rarely adequate when the model has several predictors.

Figure 15.1: Linear and Logistic Regression Models for a (0, 1) Response, for which $E(y)$ Is $P(y = 1)$

((Fig. 15.1 in 3e, replace pi on y-axis by P(y=1)))

The Logistic Regression Model for Binary Responses

Figure 15.1 also shows more realistic response curves, which have an S-shape. With these curves, the probability of a success falls between 0 and 1 for all possible x-values. These curvilinear relationships are described by the formula

$$\log\left[\frac{P(y=1)}{1-P(y=1)}\right] = \alpha + \beta x.$$

The ratio $P(y = 1)/[1 - P(y = 1)]$ equals the ***odds***, a measure Section 8.4 introduced. For instance, when $P(y = 1) = 0.75$, the odds equals $0.75/0.25 = 3.0$, meaning that a success is three times as likely as a failure. Software uses natural logarithms (base e, often denoted ln) in fitting the model. However, we won't need to use (or understand) logaritms to interpret the model and conduct inference using it.

This formula uses the log of the odds, $\log\,[P(y = 1)/(1 - P(y = 1))]$, called the ***logistic transformation***, or ***logit*** for short. The model is abbreviated as

$$\text{logit}[P(y = 1)] = \alpha + \beta x.$$

It is called the ***logistic regression model***.

When the logit follows this straight-line model, the probability $P(y = 1)$ itself follows a curve such as in Figure 15.1. The parameter β indicates whether the curve goes up or goes down as x increases. For $\beta > 0$, $P(y = 1)$ increases as x increases, as in curve (1) in Figure 15.1. For $\beta < 0, P(y = 1)$ decreases as x increases, as in curve (2) in Figure 15.1. If $\beta = 0$, $P(y = 1)$ does not change as x changes, so the curve flattens to a horizontal straight line. The steepness of the curve increases as $|\beta|$ increases. For instance, $|\beta|$ for curve (2) is greater than β for curve (1).

When $P(y = 1) = 0.50$, the odds $P(y = 1)/[1 - P(y = 1)] = 1$, and $\log[P(y = 1)/(1 - P(y = 1))] = 0$. Equating this logit value of 0 to $\alpha + \beta x$ and solving for x shows that $P(y = 1) = 0.50$ when $x = -\alpha/\beta$.

Most software uses *maximum likelihood* (see Section 5.1) to fit the model. This method is more appropriate for binary data than least squares.

Example 15.1 Income and Having Travel Credit Cards

Table 15.1 shows data for a sample of 100 adults randomly selected for an Italian study on the relation between annual income and having a travel credit card, such as American Express or Diners Club. At each level of annual income (in thousands of euros), the table indicates the number of subjects in the sample and the number of those having at least one travel credit card. Let x = annual income and y = whether have a travel credit card (1 = yes, 0 = no). For instance, for the five observations at $x = 30$, $y = 1$ for two subjects and $y = 0$ for three subjects.

Table 15.1: Annual Income (in Thousands of Euros) and Possessing a Travel Credit Card. For example, of the 5 subjects with income 30 thousand euros, 2 possessed a travel credit card.

Income	Number Cases	Credit Cards	Income	Number Cases	Credit Cards	Income	Number Cases	Credit Cards
12	1	0	21	2	0	34	3	3
13	1	0	22	1	1	35	5	3
14	8	2	24	2	0	39	1	0
15	14	2	25	10	2	40	1	0
16	9	0	26	1	0	42	1	0
17	8	2	29	1	0	47	1	0
19	5	1	30	5	2	60	6	6
20	7	0	32	6	6	65	1	1

Source: Thanks to R. Piccarreta, Bocconi Univ., Milan. The data were originally recorded in Italian lira but have been converted to euros.

Software provides results shown in Table 15.2 (heavily edited). The logistic prediction equation is

$$\text{logit}[\hat{P}(y = 1)] = -3.518 + 0.105x.$$

Since the estimate 0.105 of β is positive, the estimated probability of having a credit card increases at higher levels of income. Figure 15.2 shows the prediction curve. The estimated probability equals 0.50 at $x = -\hat{\alpha}/\hat{\beta} = (3.518)/(0.105) = 33.5$. The estimated probability of having a credit card is below 0.50 for incomes below 33.5 thousand euros and above 0.50 for incomes above this level.

Table 15.2: Fit of Logistic Regression Model for Italian Credit Card Data

	B	S.E.	Exp(B)
income	.1054	.0262	1.111
Constant	-3.5179	.7103	

Figure 15.2: Logistic Regression Prediction Curve for Example 15.1

((Fig. 15.3 in 3e, replace $\hat{\pi}$ on y-axis by $\hat{P}(y = 1)$, and change values on x-axis to 12, 24, 36, 46, 60, 72))

□

Logistic Regression Equation for Probabilities

An alternative equation for logistic regression expresses the probability of success directly. It is

$$P(y = 1) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}.$$

Here, e raised to a power represents the antilog of that number, using natural logs. (Most calculators have an e^x key that provides these antilogs, and software can report estimated probabilities based on the fit of the model.) We use this formula to estimate values of $P(y = 1)$ at particular predictor values.

From the estimates in Table 15.2, a person with annual income x thousand euros has estimated probability of having a credit card equal to

$$\hat{P}(y = 1) = \frac{e^{-3.52+0.105x}}{1 + e^{-3.52+0.105x}}.$$

For subjects with income $x = 12$, the lowest income level in this sample, the estimated probability equals

$$\hat{P}(y = 1) = \frac{e^{-3.52+0.105(12)}}{1 + e^{-3.52+0.105(12)}} = \frac{e^{-2.26}}{1 + e^{-2.26}} = \frac{0.104}{1.104} = 0.094.$$

For $x = 65$, the highest income level in this sample, the estimated probability equals 0.97.

Interpreting the Logistic Regression Model

We've seen how to estimate the probability of success, and we've seen that the sign of β tells us whether it is increasing or decreasing as x increases. How else can we interpret β? Unlike in the linear probability model, β is not the slope for the change in $P(y = 1)$ as x changes. Since the curve for $P(y = 1)$ is S-shaped, the rate at which the curve climbs or descends changes according to the value of x.

The simplest way to use β to interpret the steepness of the curve uses a straight line approximation to the logistic regression curve. A straight line drawn tangent to the curve has slope $\beta P(y = 1)[1-P(y = 1)]$, where $P(y = 1)$ is the probability at that point. Figure 15.3 illustrates. The slope is greatest when $P(y = 1) = 1/2$, where it is $\beta(1/2)(1/2) = \beta/4$. So, when $P(y = 1)$ is near 1/2, one-fourth of the β effect parameter in the logistic regression model is the approximate rate at which $P(y = 1)$ changes per 1-unit increase in x.

Figure 15.3: A Line Drawn Tangent to a Logistic Regression Curve Has Slope $\beta P(y = 1)[1 - P(y = 1)]$

((Fig. 15.2 in 3e, replace π by $P(y = 1)$))

For the Italian credit card and income data, $\hat{\beta} = 0.105$. When the estimated probability of a credit card is $\hat{P}(y = 1) = 1/2$, a line drawn tangent to the curve at that point has slope equal to $\hat{\beta}/4 = 0.105/4 = 0.026$. The rate of change in $\hat{P}(y = 1)$ for a 1 thousand euro increase in annual income is $\hat{\beta}/4 = 0.026$. At the sample mean income of $\bar{x} = 25$, the estimated probability of a credit card is $\hat{P}(y = 1) = 0.29$. Then, $\hat{\beta}\hat{P}(y = 1)[1 - \hat{P}(y = 1)] = 0.105(0.29)(0.71) = 0.022$, so a 1 thousand euro increase in annual income relates approximately to a 0.022 increase in the estimated probability of having a credit card.

Software can also fit the linear probability model, $P(y = 1) = \alpha + \beta x$. The least squares fit is $\hat{P}(y = 1) = -0.159 + 0.019x$. This formula also has about a 0.02 increase in $\hat{P}(y = 1)$ per thousand dollar increase in income. However, it provides quite different predictions at the low end and high end of the income scale. For instance, it provides the absurd prediction $\hat{P}(y = 1) > 1$ when $x \geq 84$.

Another way to describe the effect of x compares $\hat{P}(y = 1)$ at two different values of x. We've seen that when x increases from its smallest to its largest value in the sample, $\hat{P}(y = 1)$ increases from 0.09 to 0.97. This is a strong effect, as there is a very large change in $\hat{P}(y = 1)$. An alternative is to instead evaluate $\hat{P}(y = 1)$ at values of x that are less affected by outliers, such as the upper and lower quartiles.

Interpretation Using the Odds and Odds Ratio

Another interpretion of the logistic regression parameter β uses the *odds ratio* measure of association (Section 8.4). Applying antilogs to both sides of the logistic regression equation $\log\,[P(y = 1)/(1 - P(y = 1))] = \alpha + \beta x$ yields

$$\frac{P(y=1)}{1-P(y=1)} = e^{\alpha+\beta x} = e^{\alpha}(e^{\beta})^{x}.$$

The right-hand side of this equation has the *exponential regression* form studied in Section 14.6, a constant multiplied by another constant raised to the x power. This exponential relationship implies that every unit increase in x has a multiplicative effect of e^{β} on the odds.

In Example 15.1, the antilog of $\hat{\beta}$ is $e^{\hat{\beta}} = e^{0.105} = 1.11$. When annual income increases by 1 thousand euros, the estimated odds of owning a credit card multiply by 1.11; that is, they increase by 11%. When $x = 25$, for example, the estimated odds of owning a travel credit card are 1.11 times what they are when $x = 24$. When $x = 25$,

$$\text{Estimated odds} = \frac{\hat{P}(y=1)}{1-\hat{P}(y=1)} = e^{-3.518+0.105(25)} = 0.414,$$

whereas when $x = 26$,

$$\text{Estimated odds} = \frac{\hat{P}(y=1)}{1-\hat{P}(y=1)} = e^{-3.518+0.105(26)} = 0.460,$$

which is 1.11 times the value of 0.414 at $x = 25$. In other words, $e^{0.105} = 1.11 = 0.460/0.414$ is an estimated *odds ratio*, equaling the estimated odds at $x = 26$ divided by the estimated odds at $x = 25$.

Odds ratios also apply to changes in x other than 1. For example, a 10-unit change in x corresponds to a change of 10β in the log odds, and a multiplicative effect of $e^{10\beta} = (e^{\beta})^{10}$ on the odds. When $x = 30$, the odds equal $(1.11)^{10} = 2.9$ times the odds when $x = 20$.

Most software can report odds ratio estimates and estimated probabilities. Table 15.2 reports the estimated odds ratio for a 1-unit increase in x under the heading *Exp(B)*.

15.2 Multiple Logistic Regression

Logistic regression can handle multiple predictors. The multiple logistic regression model has the form

$$\text{logit}[P(y=1)] = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k.$$

The formula for the probability itself is

$$P(y=1) = \frac{e^{\alpha+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\alpha+\beta_1 x_1+\cdots+\beta_k x_k}}.$$

Exponentiating a beta parameter provides the multiplicative effect of that predictor on the odds, controlling for the other variables. The farther a β_i falls from 0, the stronger the effect of the predictor x_i, in the sense that the odds ratio falls farther from 1.

As in ordinary regression, cross-product terms allow interactions between pairs of explanatory variables. Square terms allow probabilities to increase and then decrease (or the reverse) as a predictor increases. To include categorical explanatory variables, set up dummy variables, as the next example illustrates.

Example 15.2 Death Penalty and Racial Predictors

Table 15.3 is a three-dimensional contingency table from a study[1] of the effects of racial characteristics on whether individuals convicted of homicide receive the death penalty. The variables in Table 15.3 are *death penalty verdict*, the response variable, having categories (yes, no), and the explanatory variables *race of defendant* and *race of victims*, each having categories (white, black). The 674 subjects were defendants in indictments involving cases with multiple murders in Florida between 1976 and 1987.

Table 15.3: Death Penalty Verdict by Defendant's Race and Victims' Race, for Cases with Multiple Murders in Florida

Defendant's Race	Victims' Race	Death Penalty Yes	Death Penalty No	Percent Yes
White	White	53	414	11.3
	Black	0	16	0.0
Black	White	11	37	22.9
	Black	4	139	2.8

[1] M. L. Radelet and G. L. Pierce, *Florida Law Review*, Vol. 43, 1991, pp. 1–34

For each of the four combinations of defendant's race and victims' race, Table 15.3 also lists the percentage of defendants who received the death penalty. For white defendants, the death penalty was imposed 11.3% of the time when the victims were white and 0.0% of the time when the victims were black, a difference of 11.3% − 0.0% = 11.3%. For black defendants, the death penalty was imposed 22.9% − 2.8% = 20.1% more often when the victims were white than when the victims were black. Thus, controlling for defendant's race by keeping it fixed, the percentage of *yes* death penalty verdicts was considerably higher when the victims were white than when they were black.

Now, consider the association between defendant's race and the death penalty verdict, controlling for victims' race. When the victims were white, the death penalty was imposed 22.9% − 11.2% = 11.7% more often when the defendant was black than when the defendant was white. When the victims were black, the death penalty was imposed 2.8% more often when the defendant was black than when the defendant was white. In summary, controlling for victims' race, black defendants were somewhat more likely than white defendants to receive the death penalty.

For y = death penalty verdict, let $y = 1$ denote the *yes* verdict. Since defendant's race and victims' race each have two categories, a single dummy variable can represent each. Let d be a dummy variable for defendant's race and v a dummy variable for victims' race, where

$$d = 1, \text{ defendant} = \text{white}, \quad d = 0, \text{ defendant} = \text{black},$$
$$v = 1, \text{ victims} = \text{white}, \quad v = 0, \text{ victims} = \text{black}.$$

The logistic model with main effects for these predictors is

$$\text{logit}[P(y = 1)] = \alpha + \beta_1 d + \beta_2 v,$$

where β_1 represents the effect of defendant's race, controlling for victims' race, and β_2 represents the effect of victims' race, controlling for defendant's race. Here, e^{β_1} is the odds ratio between the response variable and defendant's race, controlling for victims' race, and e^{β_2} is the odds ratio between the response and victims' race, controlling for defendant's race.

Table 15.4 shows output for the model fit (based on using the *generalized linear models* option from the *Analyze* menu in SPSS). The prediction equation is

$$\text{logit}[\hat{P}(y = 1)] = -3.596 - 0.868d + 2.404v.$$

Since $d = 1$ for white defendants, the *negative* coefficient of d means that the estimated odds of receiving the death penalty are *lower* for white defendants than for black defendants. Since $v = 1$ for white victims, the *positive* coefficient of v means that the estimated odds of receiving the death penalty are *higher* when the victims were white than when they were black.

Table 15.4: Parameter Estimates for Logistic Model for Death Penalty Data. Estimates equal 0 for the 'black' category of each predictor because its dummy variable value is 0.

	B	Std. Error	Exp(B)
(Intercept)	-3.596	.5069	.027
[defendant=white]	-.868	.3671	.420
[defendant=black]	0	.	
[victim=white]	2.404	.6006	11.072
[victim=black]	0	.	

The formula for the estimated probability of the death penalty is

$$\hat{P}(y=1) = \frac{e^{-3.596-0.868d+2.404v}}{1+e^{-3.596-0.868d+2.404v}}.$$

For instance, when the defendant is black and the victims were white, $d = 0$ and $v = 1$, so

$$\hat{P}(y=1) = \frac{e^{-3.596-0.868(0)+2.404(1)}}{1+e^{-3.596-0.868(0)+2.404(1)}} = \frac{e^{-1.192}}{1+e^{-1.192}} = \frac{0.304}{1.304} = 0.233.$$

This is close to the sample proportion of 0.229 (Table 15.3). The estimated probabilities, unlike sample proportions, perfectly satisfy the model. The closer the sample proportions fall to the estimated probabilities, the better the model fits.

The antilog of $\hat{\beta}_1$, namely, $e^{\hat{\beta}_1} = e^{-0.868} = 0.42$, is the estimated odds ratio between defendant's race and the death penalty, controlling for victims' race. The estimated odds of the death penalty for a white defendant equal 0.42 times the estimated odds for a black defendant. We list *white* before *black* in this interpretation, because the dummy variable was set up with $d = 1$ for *white* defendants. If we instead let $d = 1$ for black defendants rather than white, then we get $\hat{\beta}_1 = 0.868$ instead of -0.868. Then, $e^{0.868} = 2.38$, which is $1/0.42$; that is, the estimated odds of the death penalty for a black defendant equal 2.38 times the estimated odds for a white defendant, controlling for victims' race.

For victims' race, $e^{2.404} = 11.1$. Since $v = 1$ for *white* victims, the estimated odds of the death penalty when the victims were *white* equal 11.1 times the estimated odds when the victims were black, controlling for defendant's race. This is a very strong effect.

This model assumes that both explanatory variables affect the response variable, but with a lack of interaction: The effect of defendant's race on the death penalty verdict is the same for each victims' race and the effect of victims' race is the same for each defendant's race. This means that the estimated odds ratio between each predictor and the response takes the same value at each category of the other predictor. For instance, the estimated odds ratio of 11.1 between victims' race and the death penalty is the same when the defendants were white as when the defendants were black.

□

Effects on Odds

The parameter estimates for the logistic regression model are linear effects, but on the scale of the *log* of the odds. It is easier to understand effects on the *odds* scale than the *log odds* scale. The antilogs of the parameter estimates are multiplicative effects on the odds.

To illustrate, for the data on the death penalty, the prediction equation

$$\log[\hat{P}(y=1)/(1-\hat{P}(y=1))] = -3.596 - 0.868d + 2.404v$$

refers to the log odds (i.e., logit). The corresponding prediction equation for the odds is

$$\text{odds} = e^{-3.596-0.868d+2.404v} = e^{-3.596}e^{-0.868d}e^{2.404v}.$$

For white defendants, $d = 1$, and the estimated odds equal $e^{-3.596}e^{-0.868}e^{2.404v}$. For black defendants, $d = 0$, and the estimated odds equal $e^{-3.596}e^{2.404v}$. The estimated odds for white defendants divided by the estimated odds for black defendants equal $e^{-0.868} = 0.42$. This shows why the antilog of the coefficient for d in the prediction equation is the estimated odds ratio between defendant's race and death penalty verdict, for each fixed victims' race. The effect of defendant's race being white is to multiply the estimated odds of a *yes* death penalty verdict by $e^{-0.868} = 0.42$ compared to its value for black defendants. The actual values of the odds depend on victims' race, but the ratio of the odds is the same for each.

The logit model expression for the log odds is *additive*, but taking antilogs yields a *multiplicative* expression for the odds. In other words, the antilogs of the parameters are *multiplied* to obtain odds. We can use this expression to calculate odds estimates for any combination of defendant's race and victims' race. For instance, when the defendant is black ($d = 0$) and the victims were white ($v = 1$), the estimated odds of the death penalty are

$$\text{odds} = e^{-3.596-0.868d+2.404v} = e^{-3.596-0.868(0)+2.404(1)} = e^{-1.192} = 0.304.$$

Since $P(y = 1) = \text{odds}/(1 + \text{odds})$, the ratio $0.304/(1 + 0.304) = 0.233$ is the estimated probability of the death penalty found above.

Example 15.3 Factors Affecting First Home Purchase

Table 15.5 summarizes logistic regression results from a study[2] of how family transitions relate to first home purchase by young married households. The response variable is whether the subject owns a home (1 = yes, 0 = no). The parameter $P(y = 1)$ is the probability of home ownership.

Table 15.5: Results of Logistic Regression for the Probability of Home Ownership

Variable	Estimate	Std. Error
Intercept	-2.870	—
Husband earnings ($10,000)	0.569	0.088
Wife earnings ($10,000)	0.306	0.140
N0. years married	-0.039	0.042
Married in 2 years (1 = yes)	0.224	0.304
Working wife in 2 years (1 = yes)	0.373	0.283
No. children	0.220	0.101
Add child in 2 years (1 = yes)	0.271	0.140
Head's education (no. years)	-0.027	0.032
Parents' home ownership (1 = yes)	0.387	0.176

The model contains several explanatory variables. Two variables measure current income: husband's earnings and wife's earnings. Two variables measure marital duration: the number of years the respondent has been married and marital status two years after the year of observation. The latter is a categorical variable with levels (married, married with a working wife, single), single being the omitted category for the two dummy variables. Two variables measure presence of children: the number of children aged 0–17 in the household and a dummy variable for whether the family has more children aged 0–17 two years after the year of observation. The other variables are the household head's education and a dummy variable for whether the subject's parents owned a home in the last year that the subject lived in the parental home.

In Table 15.5, the ratio of the estimate to its standard error exceeds 2 for four of the predictors. Other things being fixed, the probability of home ownership increases with husband's earnings, wife's earnings, the number of children, and parents' home ownership. For instance,

[2] from J. Henretta, *Social Forces*, Vol. 66, 1987, pp. 520–536

each additional child had the effect of multiplying the estimated odds of owning a home by $e^{0.220} = 1.25$; that is, the estimated odds increase by 25%. A \$10,000 increase in earnings had the effect of multiplying the estimated odds of owning a home by $e^{0.569} = 1.77$ if the earnings add to husband's income and by $e^{0.306} = 1.36$ for wife's income.

From Table 15.5, number of years married, marital status in two years, and head's education show little evidence of an effect, given the other variables in the model. We could refit the model without these predictors. This more parsimonious model may yield better estimates of the effects of the significant variables.

□

Effects on Probabilities

We've seen that we can summarize the effects of predictors by estimating *odds ratios*. Many researchers find it easier to get a feel for the effects by viewing summaries that use the *probability* scale. Such summaries can report estimated probabilities at particular values of a predictor of interest. This evaluation is done at fixed values of the other predictors, such as at their means or at certain values of interest.

To illustrate, let's investigate the effect of husband's earnings for the above example. Consider the estimated probability $\hat{P}(y = 1)$ of home ownership when wife's earnings = \$50,000, years married = 3, the wife is working in two years, number of children = 0, add child in two years = 0, head's education = 16 years, and parents' home ownership = 0. When husband's earnings = \$20,000,

$$\hat{P}(y = 1) = \frac{e^{-2.870+0.569(2)+0.306(5)-0.039(3)+0.373(1)-0.027(16)}}{1 + e^{-2.870+0.569(2)+0.306(5)-0.039(3)+0.373(1)-0.027(16)}} = 0.41.$$

Then, $\hat{P}(y = 1)$ increases to 0.55 if husband's earnings increase to \$30,000, to 0.79 if they increase to \$50,000, and to 0.98 if they increase to \$100,000. The effect seems quite strong.

Alternatively, you can report the change in the estimated probability $\hat{P}(y = 1)$ when a predictor increases by a certain amount, such as (1) by a fixed value (e.g., 1), (2) by a standard deviation, (3) over its range from its lowest to greatest value, or (4) over its inter-quartile range from the lower quartile to the upper quartile. Approach (4) is, unlike (1), not affected by the choice of scale and, unlike (2) and (3), not affected by outliers.

15.3 Inference for Logistic Regression Models

We next discuss statistical inference for logistic regression. As usual, inference assumes randomization for gathering the data. It also assumes a *binomial* distribution for the response variable. The model identifies y as having a binomial distribution and uses the logit link function for $P(y = 1)$, which is the mean of y.

Wald and Likelihood-Ratio Tests of Independence

For the bivariate logistic regression model

$$\text{logit}[P(y = 1)] = \alpha + \beta x,$$

H_0: $\beta = 0$ states that x has no effect on $P(y = 1)$. This is the *independence* hypothesis. Except for very small samples, we can test H_0 using a z test statistic, dividing the maximum likelihood estimate $\hat{\beta}$ by its standard error. Some software reports the square of this statistic, called a ***Wald statistic***. This has a chi-squared distribution with $df = 1$. It has the same P-value as the z statistic for the two-sided H_a: $\beta \neq 0$.

Most software can also report another test for this hypothesis. Called the ***likelihood-ratio test***, in general it is a way to compare two models, a full model and a simpler model. It tests that the extra parameters in the full model equal zero. For example, for bivariate logistic regression it tests H_0: $\beta = 0$ by comparing the model $\text{logit}[P(y = 1)] = \alpha + \beta x$ to the simpler model $\text{logit}[P(y = 1)] = \alpha$. The test uses a key ingredient of maximum likelihood inference, the *likelihood function*. Denoted by ℓ, this gives the probability of the observed data as a function of the parameter values. The maximum likelihood estimates maximize this function. (Specifically, the estimates are the parameter values for which the observed data are most likely; see Section 5.1).

Let ℓ_0 denote the maximum of the likelihood function when H_0 is true, and let ℓ_1 denote the maximum without that assumption. The formula for the likelihood-ratio test statistic is

$$-2\log\left(\frac{\ell_0}{\ell_1}\right) = (-2\log\ell_0) - (-2\log\ell_1)$$

It compares the maximized values of $(-2\log\ell)$ when H_0 is true and when it need not be true. There is a technical reason for using -2 times the log of this ratio, namely, that the test statistic then has approximately a chi-squared distribution for large samples. The df value equals the number of parameters in the null hypothesis.

For testing H_0: $\beta = 0$ with large samples, the Wald test and likelihood-ratio test usually provide similar results. For small to moderate sample sizes, the likelihood-ratio statistic often tends to be larger and gives a more powerful test than the Wald statistic. To be safe, use it rather than the Wald statistic.

Example 15.4 Inference for Income and Travel Credit Cards

Table 15.6 shows inference results for Example 15.1 (page 690) about how the probability of having a travel credit card depends on income level. H_0: $\beta = 0$ states that the probability of having a travel credit card is the same at all income levels.

Table 15.6: Logistic Regression Inference for Italian Credit Card Data

	B	S.E.	Wald	df	Sig.	95% CI for exp(B)	
income	.1054	.0262	16.24	1	.000	1.056	1.170
Constant	-3.5179	.7103	24.53	1	.000		

From the table, $\hat{\beta} = 0.1054$ has an estimated standard error of 0.0262. The z test statistic equals $z = 0.1054/0.0262 = 4.02$. This has a P-value of 0.000 for H_a: $\beta \neq 0$. Table 15.6 reports the square of this statistic, the Wald statistic. This chi-squared statistic equals 16.2, with $df = 1$, and has the same P-value. There is strong evidence of a positive association between income and having a travel credit card.

The likelihood-ratio test of H_0: $\beta = 0$ compares the maximized value of $(-2 \log \ell)$ for this model, $\text{logit}[P(y = 1)] = \alpha + \beta x$, to its maximized value for the simpler model with $\beta = 0$. In its Model Summary, SPSS reports a $-2 \log \ell$ value of 97.23 for the model. Compared to the model with only an intercept term, SPSS reports (e.g., if the model is built with the "Forward:LR" method) that the change in $-2 \log \ell$ would be 26.59 with $df = 1$. This is the likelihood-ratio test statistic. It also has a P-value of 0.000.

The result of these tests is no surprise. We would expect subjects with higher incomes to be more likely to have travel credit cards. As usual, confidence intervals are more informative. We could obtain confidence intervals for β or for $P(y = 1)$ at various values of x. The formula for a confidence interval for $P(y = 1)$ is beyond the scope of this text, but is simple with some software (e.g., SAS, as the appendix shows). A 95% confidence interval for the probability of having a credit card equals (0.04, 0.20) at the lowest income level of 12, (0.20, 0.40) at the mean income level of 25, and (0.78, 0.996) at the highest income level of 65.

□

Inference in Multiple Logistic Regression

For testing the partial effect of a predictor in a multiple logistic regression model, the parameter estimate divided by its standard error is a z test statistic. The square of that, the Wald statistic, is a chi-squared statistic with $df = 1$. Most software also reports likelihood-ratio tests, which compare the $(-2\log \ell)$ values with and without the predictor in the model. This is particularly useful if a categorical predictor has several levels, in which case it has several dummy variables and the test of its effect equates several parameters to 0 in H_0.

Example 15.5 Inference for Death Penalty and Racial Predictors

For the death penalty data, Example 15.2 (page 694) used the model

$$\text{logit}[P(y=1)] = \alpha + \beta_1 d + \beta_2 v,$$

with dummy variables d and v for defendant's race and victims' race. If $\beta_1 = 0$, the death penalty verdict is independent of defendant's race, controlling for victims' race. In that case, the odds ratio between the death penalty verdict and defendant's race equals $e^0 = 1$ for each victims' race.

Software (SPSS) shows the results

	B	Std. Error	Wald Chi-Sq.	Sig.	95% CI for Exp(B)	
(Intercept)	-3.596	.5069	50.33	.000		
[defendant=white]	-.868	.3671	5.59	.018	.20	.86
[victim=white]	2.404	.6006	16.03	.000	3.41	35.93

The defendant's race effect of -0.868 has a standard error of 0.367. The z test statistic for H_0: $\beta_1 = 0$ is $z = -0.868/0.367 = -2.36$, and the Wald statistic equals $(-2.36)^2 = 5.59$ (shown in the table). For the two-sided alternative, either statistic has a P-value of 0.018. Similarly, the test of H_0: $\beta_2 = 0$ has $P = 0.000$ and provides extremely strong evidence of a victims' race effect.

The parameter estimates are also the basis of confidence intervals for odds ratios. Since the estimates refer to *log* odds ratios, after constructing the interval for a β_j we take antilogs of the endpoints to form the interval for the odds ratio. For instance, since the estimated log odds ratio of 2.404 between victims' race and the death penalty has a standard error of 0.601, a 95% confidence interval for the true log odds ratio is

$$2.404 \pm 1.96(0.601), \quad \text{or } (1.23, 3.58).$$

From applying the antilog (i.e., exponential function, e^x) to each endpoint, the confidence interval for the odds ratio equals $(e^{1.23}, e^{3.58}) =$

(3.4, 35.9), shown in the table. For a given defendant's race, when the victims were white, the estimated odds of the death penalty are between 3.4 and 35.9 times the estimated odds when the victims were black.

Most software can also provide confidence intervals for probabilities. 95% confidence intervals for the probability of the death penalty are (0.14, 0.37) for black defendants with white victims, (0.01, 0.07) for black defendants with black victims, (0.09, 0.15) for white defendants with white victims, and (0.003, 0.04) for white defendants with black victims.

□

Likelihood-Ratio Test Comparing Logistic Regression Models

To compare a model with a set of predictors to a simpler model having fewer predictors, the likelihood-ratio test uses the difference in the values of $(-2\log\ell)$ for the two models. This is an approximate chi-squared statistic with *df* given by the number of extra parameters in the full model. This test is an analog of the F test for comparing complete and reduced regression models (Section 11.6).

To illustrate, the model in Example 15.3 about home ownership (page 698) has $(-2\log\ell) = 2931.2$. After adding five variables to the model relating to the housing market, such as the median sale price of existing homes in the area, $-2\log\ell$ drops to 2841.1. The difference $(2931.2 - 2841.1) = 90.1$ is a chi-squared statistic with $df = 5$, since the more complex model has five additional parameters. This shows extremely strong evidence of a better fit for the more complex model ($P < 0.0001$). So, at least one of these variables provides an improvement in predictive power.

A comparison of the $(-2\log\ell)$ values for a model and for the model containing only an intercept term tests the joint effects of *all* the predictors. This test is a chi-squared analog for binary data of the F test in regression of $H_0 : \beta_1 = \cdots = \beta_k = 0$ that *none* of the predictors have an effect on y.

15.4 Logistic Regression Models for Ordinal Variables*

Many applications have a categorical response variable with more than two categories. For instance, the General Social Survey recently asked subjects whether government spending on the environment should increase, remain the same, or decrease. An extension of logistic regression can handle ordinal response variables.

Cumulative Probabilities and their Logits

Let y denote an ordinal response variable. Let $P(y \leq j)$ denote the probability that the response falls in category j or below (i.e., in category 1, 2, ..., or j). This is called a ***cumulative probability.*** With four categories, for example, the cumulative probabilities are

$$P(y = 1), P(y \leq 2) = P(y = 1)+P(y = 2), P(y \leq 3) = P(y = 1)+P(y = 2)+P(y = 3),$$

and the final cumulative probability uses the entire scale, so $P(y \leq 4) = 1$.

A c-category response has c cumulative probabilities. The order of forming the cumulative probabilities reflects the ordering of the response scale. The probabilities satisfy

$$P(y \leq 1) \leq P(y \leq 2) \leq \cdots \leq P(y \leq c) = 1.$$

The odds of response in category j or below is the ratio

$$\frac{P(y \leq j)}{P(y > j)}.$$

For instance, when the odds equal 2.5, the probability of response in category j or below equals 2.5 times the probability of response above category j. Each cumulative probability can convert to an odds.

A popular logistic model for an ordinal response uses logits of the cumulative probabilities. With $c = 4$, for example, the logits are

$$\text{logit}[P(y \leq 1)] = \log\left[\frac{P(y = 1)}{P(y > 1)}\right] = \log\left[\frac{P(y = 1)}{P(y = 2) + P(y = 3) + P(y = 4)}\right]$$

$$\text{logit}[P(y \leq 2)] = \log\left[\frac{P(y \leq 2)}{P(y > 2)}\right] = \log\left[\frac{P(y = 1) + P(y = 2)}{P(y = 3) + P(y = 4)}\right]$$

$$\text{logit}[P(y \leq 3)] = \log\left[\frac{P(y \leq 3)}{P(y > 3)}\right] = \log\left[\frac{P(y = 1) + P(y = 2) + P(y = 3)}{P(y = 4)}\right].$$

Since the final cumulative probability necessarily equals 1.0, we exclude it from the model. These logits of cumulative probabilities are called ***cumulative logits.*** Each cumulative logit regards the response as binary by considering whether the response is at the low end or the high end of the scale, where "low" and "high" have a different definition for each cumulative logit.

Cumulative Logit Models for an Ordinal Response

A model can simultaneously describe the effect of an explanatory variable on all the cumulative probabilities for y. For each cumulative probability, the model looks like an ordinary logistic regression, where the two outcomes are low = "category j or below" and high = "above category j." This model is

$$\text{logit}[P(y \leq j)] = \alpha_j - \beta x, \quad j = 1, 2, \ldots, c-1.$$

For $c = 4$, for instance, this single model describes three relationships: the effect of x on the odds that $y \leq 1$ instead of $y > 1$, the effect of x on the odds that $y \leq 2$ instead of $y > 2$, and the effect of x on the odds that $y \leq 3$ instead of $y > 3$. The model requires a separate intercept parameter α_j for each cumulative probability. Since the cumulative probabilities increase as j increases, so do $\{\alpha_j\}$.

Why is the model written with a minus sign before β? This is not necessary, but it is how the model is parameterized by some software, such as SPSS. That way, when $\beta > 0$, when x is *higher* cumulative probabilities are *lower*. But cumulative probabilities being lower means it is less likely to observe relatively low values and thus more likely to observe *higher* values of y. So, this parameterization accords with the usual formulation of a positive association, in the sense that a positive β corresponds to a positive association (higher x tending to occur with higher y). Statistical software for fitting the model has no standard convention. Software (such as SAS) that specifies the model as $\text{logit}[P(y \leq j)] = \alpha_j + \beta x$ will report the opposite sign for $\hat{\beta}$. You should be careful to check how your software defines the model so you interpret $\hat{\beta}$ properly.

The parameter of main interest, β, describes the effect of x on y. When $\beta = 0$, each cumulative probability does not change as x changes, and the variables are independent. The effect of x increases as $|\beta|$ increases. In this model, β does not have a j subscript. It has the same value for each cumulative logit. In other words, the model assumes that the effect of x is the same for each cumulative probability. This cumulative logit model with this common effect is often called the ***proportional odds*** model.

Figure 15.4 depicts the model for four response categories with a quantitative predictor. The model implies a separate S-shaped curve for each of the three cumulative probabilities. For example, the curve for $P(y \leq 2)$ has the appearance of a logistic regression curve for a binary response with pair of outcomes $(y \leq 2)$ and $(y > 2)$. At any fixed x value, the three curves have the same ordering as the cumulative probabilities, the one for $P(y \leq 1)$ being lowest.

The size of $|\beta|$ determines how quickly the curves climb or drop. The common value for β means that the three response curves have

Figure 15.4: Depiction of Curves for Cumulative Probabilities in Cumulative Logit Model

((Fig. 15.4 in 3e))

the same shape. In Figure 15.4, the curve for $P(y \leq 1)$ is the curve for $P(y \leq 2)$ moved to the right and the curve for $P(y \leq 3)$ moved even further to the right. To describe the association, we can form $e^{-\beta}$ as a multiplicative effect of a 1-unit increase in x on the odds for the cumulative probabilities: For each j, the odds that $y \leq j$ multiply by $e^{-\beta}$ for each one-unit increase in x. (It is $e^{-\beta}$ rather than e^{β} because of the negative sign attached to β in the model formula.)

Model fitting treats the observations as independent from a ***multinomial distribution***. This is a generalization of the *binomial distribution* from two to multiple outcome categories. Software estimates the parameters using all the cumulative probabilities at once. This provides a single estimate $\hat{\beta}$ for the effect of x, rather than the three separate estimates we'd get by fitting the model separately for each cumulative probability. If you reverse the order of categories of y (i.e., listing from high to low instead of from low to high), the model fit is the same but the sign of $\hat{\beta}$ reverses.

Example 15.6 Comparing Political Ideology of Democrats and Republicans

Do Republicans tend to be more conservative than Democrats? Table 15.7, for the subjects of age 18-30 in the 2004 General Social Survey, relates political ideology to party affiliation. We treat political ideology, which has a five-point ordinal scale, as the response variable. In Chapter 12 on ANOVA methods, we treated political ideology as quantitative by assigning scores to categories. Now, we treat it as ordinal by using a cumulative logit model.

Let x be a dummy variable for party affiliation, with $x = 1$ for Democrats and $x = 0$ for Republicans. Table 15.8 shows results of fitting the cumulative logit model. The table reports four intercept parameter estimates, because the response variable (political ideology) has five categories. These estimates are not as relevant as the estimated effect of the explanatory variable (party affiliation), which is $\hat{\beta} = -2.241$. Since the dummy variable x is 1 for Democrats and since high values of y represent greater conservatism, the *negative* $\hat{\beta}$ value means that Democrats tend to be *less* conservative than Republicans. Democrat are more likely than

Table 15.7: Political Ideology by Party Affiliation

	Political Ideology				
Party Affiliation	Very Liberal	Slightly Liberal	Moderate	Slightly Conservative	Very Conservative
Democratic	29	17	36	4	5
Republican	2	7	23	23	19

Republicans to fall toward the liberal end of the political ideology scale.

Table 15.8: Printout for Cumulative Logit Model Fitted to Table 15.7

```
                          Estimate  Std. Error   Wald    df    Sig.
INTERCP1 [ideology=1]      -3.051      .336      82.30    1    .000
INTERCP2 [ideology=2]      -2.150      .300      51.28    1    .000
INTERCP3 [ideology=3]       -.183      .226        .66    1    .418
INTERCP4 [ideology=4]        .983      .250      15.41    1    .000
PARTY                      -2.241      .338      44.05    1    .000
```

We can also interpret $\hat{\beta} = -2.241$ by exponentiating $-\hat{\beta}$ to form an estimated odds ratio for cumulative probabilities. For any fixed j, the estimated odds that a Democrat's response is in the liberal direction rather than the conservative direction (i.e., $y \leq j$ rather than $y > j$) equal $e^{-\hat{\beta}} = e^{2.241} = 9.4$ times the estimated odds for Republicans. Specifically, this odds ratio applies to each of the four cumulative probabilities for Table 15.7.

To illustrate, using the second of the four cumulative probabilities, the estimated odds that a Democrat is very liberal or slightly liberal, rather than moderate, slightly conservative, or very conservative, are 9.4 times the estimated odds that a Republican is very liberal or slightly liberal, rather than moderate, slightly conservative, or very conservative. The value 9.4 is far from the no effect value of 1.0. The sample has a strong association, Democrats tending to be more liberal than Republicans.

□

Inference for Effects on an Ordinal Response

When $\beta = 0$ in the cumulative logit model, the variables are independent. We test independence by testing H_0: $\beta = 0$. As usual, the z test statistic divides $\hat{\beta}$ by its standard error. The square of that ratio is the Wald statistic, which is chi-squared with $df = 1$. Better yet, the likelihood-ratio test is based on the difference in $(-2 \log \ell)$ values with and without the predictor in the model. Most software can report both these tests.

From Table 15.8, the effect of party affiliation has estimate $\hat{\beta} = -2.241$ and standard error $= 0.338$. So, $z = -2.241/0.338 = 6.63$ and the Wald statistic equals $(6.63)^2 = 44.0$. Not reported here is the likelihood-ratio test statistic for this hypothesis, which equals 51.8, based on $df = 1$. This is the difference between values of $(-2 \log \ell)$ for the model with only the intercept terms (507.8) and the full model (456.1). With either statistic, the P-value is 0.000 for testing H_0: $\beta = 0$ (independence of political ideology and party) against H_a: $\beta \neq 0$. These tests and $\hat{\beta}$ provide strong evidence that Democrats are politically more liberal than Republicans.

This test of independence takes into account the ordering of the response categories. It is usually more powerful than tests of independence that ignore the ordering, such as the Pearson chi-squared test of Section 8.2. When there truly is dependence, the ordinal test usually yields a smaller P-value.

A confidence interval for the odds ratio describes the association for the cumulative logit model. Since $\hat{\beta} = -2.241$ with $se = 0.338$, the 95% confidence interval for the population log odds ratio represented by $-\beta$ equals $2.241 \pm 1.96(0.338)$, or (1.58, 2.90). The confidence interval for the odds ratio is $(e^{1.58}, e^{2.90})$, or (4.9, 18.2). The odds that a Democrat's response falls in the liberal direction are more than about five times the odds for Republicans. To illustrate, using the second of the four cumulative probabilities, we conclude that the odds that a Democrat is very liberal or slightly liberal, rather than moderate, slightly conservative, or very conservative, fall between 4.9 and 18.2 times the odds for a Republican.

Invariance to Choice of Response Categories

When the cumulative logit model fits well, it also fits well with similar effects for any collapsing of the response categories. For instance, if a model for categories (Very liberal, Slightly liberal, Moderate, Slightly conservative, Very conservative) fits well, approximately the same estimated effects result when we fit the model to the data after collapsing the response scale to (Liberal, Moderate, Conservative). This *invariance* to the choice of response categories is a nice feature of the model. Two researchers who use different response categories in studying an association should reach similar conclusions.

To illustrate, we collapse Table 15.7 to a three-category response, combining the two liberal categories and combining the two conservative categories. The estimated effect of party affiliation changes only from

-2.241 (standard error = 0.338) to -2.258 (standard error = 0.343). Some loss of efficiency occurs in collapsing ordinal scales, resulting in larger standard errors. In practice, when observations are spread fairly evenly among the categories, the efficiency loss is minor unless the collapsing is to a binary response. It is usually inadvisable to collapse ordinal data to binary.

The cumulative logit model implies trends upward or downward among distributions of y at different values of explanatory variables. When x refers to two groups, as in Table 15.7, the model fits well when subjects in one group tend to make higher responses on the ordinal scale than subjects in the other group. The model does not fit well when the response distributions differ in their *variability* rather than their average. If Democrats tended to be primarily moderate in political ideology, while Republicans tended to be both very conservative and very liberal (i.e., at the two extremes of the scale), then the Republicans' responses would show greater variability than the Democrats'. The two political ideology distributions would be quite different, but the model would not detect this if the average responses were similar.

Extensions of the Model for Multiple Predictors

Cumulative logit models can handle multiple explanatory variables, which can be quantitative and/or categorical. When an explanatory variable is categorical, dummy variables represent the categories.

To check the fit of the model, we can analyze whether extra terms, such as interactions, provide a significant improvement in the model fit. One way to do this uses a likelihood-ratio test of whether the extra parameters equal 0. Some software also provides a chi-squared test for the *proportional odds* assumption that the β effects are the same for all cumulative probabilities.

As in other statistical endeavors, there is danger in putting too much emphasis on statistical tests, whether of effects or of goodness of fit. Results are sensitive to sample size, more significant results tending to occur with larger sample sizes. Test statistics merely indicate the level of parsimony that is possible. It is important to supplement the tests with estimation methods that describe the strength of effects and with residual analyses that detect parts of the data for which the overall trend fails to hold.

Ordinal Predictor in a Logistic Regression Model

This section has focused on modeling ordinal response variables. It is usually advisable to treat ordinal explanatory variables in a quantitative manner, for any type of regression model.

To illustrate, for Table 15.7, let's now treat party affiliation as the response variable and political ideology as the explanatory variable. Let $y = 1$ if the subject identifies as a Democrat. Rather than assign dummy variables for the categories of x = political ideology, we could assign scores to the categories and assume

$$\text{logit}[P(y = 1)] = \alpha + \beta x.$$

The assignment of scores means that ideology is treated as a quantitative variable. The model is more parsimonious and simpler to interpret than if we used dummy variables.

For Table 15.7, it seems sensible to assign equally spaced scores, such as $\{1, 2, 3, 4, 5\}$. The prediction equation for the probability that party affiliation is Democrat is then

$$\text{logit}[\hat{P}(y = 1)] = 3.404 - 1.062x.$$

For this prediction equation, $e^{-1.062} = 0.35$ is the multiplicative effect of a single category increase in conservatism on the odds of being a Democrat instead of Republican. The odds of being classified Democrat multiply by 0.35 for every category increase in conservatism. Thus, the probability of being a Democrat decreases considerably as conservatism increases.

15.5 Logistic Models for Nominal Responses*

For nominal response variables, an extension of the binary logistic regression model provides an ordinary logistic model for each pair of response categories. The models simultaneously use all pairs of categories by specifying the odds of outcome in one category instead of another. The order of listing the categories is irrelevant, because the model treats the response scale as nominal.

15.5.1 Baseline-Category Logits

Logit models for nominal response variables pair each category with a baseline category. Most software uses the last category as the baseline. With three response categories, for example, the *baseline-category logits* are

$$\log\left[\frac{P(y = 1)}{P(y = 3)}\right], \quad \log\left[\frac{P(y = 2)}{P(y = 3)}\right]$$

For c outcome categories, the baseline-category logit model with a predictor x is

$$\log\left[\frac{P(y = j)}{P(y = c)}\right] = \alpha_j + \beta_j x, \ \ j = 1, ..., c - 1.$$

Table 15.9: Belief in Afterlife by Sex and Race

		Belief in Afterlife		
Race	Sex	Yes	Undecided	No
Black	Female	64	9	15
	Male	25	5	13
White	Female	371	49	74
	Male	250	45	71

Given that the response falls in category j or the last category, this models the log odds that the response is j. It looks like an ordinary logistic regression model, where the two outcomes are category j and category c.

Each logit has its own parameters. Software for multicategory logit models fits all the equations *simultaneously*, again assuming independent observations from multinomial distributions.

Example 15.7 Belief in Afterlife by Sex and Race

Table 15.9, from a General Social Survey, has y = belief in life after death, with categories (Yes, Undecided, No), and explanatory variables sex and race. Let $s = 1$ for females and 0 for males, and $r = 1$ for blacks and 0 for whites. With 'no' as the baseline category for y, the model is

$$\log\left[\frac{P(y=1)}{P(y=3)}\right] = \alpha_1 + \beta_1^S s + \beta_1^R r,$$

$$\log\left[\frac{P(y=2)}{P(y=3)}\right] = \alpha_2 + \beta_2^S s + \beta_2^R r.$$

The S and R superscripts identify the sex and race parameters.

The model assumes a lack of interaction between sex and race in their effects on belief in life after death. Table 15.10 shows the parameter estimates. The first equation refers to the odds of a 'yes' rather than a 'no' response on belief in the afterlife. We see that $\beta_1^S = 0.419$ and $\beta_1^R = -0.342$. Since the dummy variables are 1 for females and for blacks, given that a subject's response was 'yes' or 'no,' the estimated probability of a 'yes' response was higher for females than for males (given race) and lower for blacks than for whites (given sex).

The effect parameters represent log odds ratios with the baseline category. For instance, $\beta_1^S = 0.419$ is the conditional log odds ratio between sex and response categories 1 and 3 (yes and no), given race. For females,

Table 15.10: Output for Baseline-Category Logit Model Fitted to Table 15.9. First equation uses belief categories (yes, no) and second equation uses belief categories (undecided, no). Estimates for second category of race and sex are 0 because of parameter redundancy (not having their own dummy variable).

```
------------------------------------------
Belief                    B    Std. Error
1      Intercept       1.225     .128
       [sex=female]     .419     .171
       [sex=male]         0         .
       [race=black]    -.342     .237
       [race=white]       0         .

2      Intercept       -.487     .183
       [sex=female]     .105     .247
       [sex=male]         0         .
       [race=black]    -.271     .354
       [race=white]       0         .

The reference category is 3.
------------------------------------------
```

the estimated odds of response 'yes' rather than 'no' on life after death are $e^{0.419} = 1.52$ times those for males, controlling for race. For blacks, the estimated odds of response 'yes' rather than 'no' on life after death are $e^{-0.342} = 0.71$ times those for whites, controlling for sex.

□

The choice of the baseline category is arbitrary. We can use the equations for a given choice to get equations for *any* pair of categories. For example, using properties of logarithms,

$$\log\left[\frac{P(y=1)}{P(y=2)}\right] = \log\left[\frac{P(y=1)}{P(y=3)}\right] - \log\left[\frac{P(y=2)}{P(y=3)}\right].$$

So, to get the prediction equation for the log odds of belief 'yes' instead of 'undecided' in the previous example, we take

$$(1.225+0.419s-0.342r)-(-0.487+0.105s-0.271r) = 1.71+0.314s-0.071r.$$

For example, the odds of response 'yes' instead of 'undecided' are higher for females (for whom $s = 1$) than males ($s = 0$).

Inference applies as in ordinary logistic regression, except now to test the effect of a predictor we consider all its parameters for the various equations. For example, in the above example, the test of the sex effect has H_0: $\beta_1^S = \beta_2^S = 0$. The likelihood-ratio test compares the model to the simpler model dropping sex as a predictor. The test statistic equals 7.2. It has $df = 2$, since H_0 has 2 parameters. The P-value of 0.03 shows evidence of a sex effect. By contrast, the effect of race is not significant, the likelihood-ratio statistic equaling 2.0 on $df = 2$. This partly reflects the larger standard errors that the effects of race have, due to a much greater imbalance between sample sizes in the race categories than in the sex categories.

15.6 Loglinear Models for Categorical Variables*

Logistic regression models are similar in structure to ordinary regression models, both types predicting a response variable using explanatory variables. By contrast, ***loglinear models*** are appropriate for contingency tables in which each classification is a response variable. Loglinear analysis resembles a correlation analysis more than a regression analysis. The loglinear focus is on studying associations between pairs of variables rather than modeling the response on one of them in terms of the others.

Loglinear models are special cases of generalized linear models that assume *Poisson* distributions for cell counts in contingency tables. The Poisson distribution is defined for discrete variables, such as counts, that can take nonnegative integer values. Equivalently, given the overall sample size, they assume a *multinomial* distribution for the counts in the cells of the contingency table.

The loglinear model formulas express the logs of cell expected frequencies in terms of dummy variables for the categorical variables and interactions between those variables. The actual model formulas can be cumbersome, and this section instead uses a symbolic notation that highlights the pairs of variables that are associated. Exercise 39 shows why the models are called *loglinear* models.

Example 15.8 Students' Use of Alcohol, Cigarette, and Marijuana

Table 15.11 comes from a survey by the Wright State University School of Medicine and the United Health Services in Dayton, Ohio. The survey asked senior high-school students in a nonurban area near Dayton, Ohio, whether they had ever used alcohol, cigarettes, or marijuana. Table 15.11 is a $2 \times 2 \times 2$ contingency table that cross classifies responses on these three items.

In this table, all three variables are response variables, rather than one a response and the others explanatory. The models presented in this section describe their association structure. They analyze whether each

Table 15.11: Alcohol (A), Cigarette (C), and Marijuana (M) Use for High School Seniors

Alcohol Use	Cigarette Use	Marijuana Use Yes	Marijuana Use No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

Source: I am grateful to Prof. Harry Khamis, Wright State Univ., for these data.

pair of variables is associated and whether the association is the same at each category of the third variable.

□

A Hierarchy of Loglinear Models for Three Variables

Loglinear models apply to contingency tables with any number of dimensions. We use three-way tables to introduce basic ideas, illustrating for Table 15.11. Denote the three variables by X, Y, and Z. (In this section, to emphasize that we are considering probabilistic properties of the variables rather than their particular values, we use upper-case notation for the variables.)

Loglinear models describe ***conditional associations*** in partial tables that relate two of the variables while controlling for the third one. A pair of variables could be statistically independent at each category of the third variable. In other words, the population version of each partial table could satisfy independence. In that case, the pair is said to be ***conditionally independent***, and the odds ratios equal 1 in the partial tables. Or, associations might exist in some or all of the partial tables. We now introduce a hierarchy of five loglinear models, ordered in terms of the extent of association.

1. All three pairs of variables are conditionally independent. That is,
   X is independent of Y, controlling for Z;
   X is independent of Z, controlling for Y;
   Y is independent of Z, controlling for X.

2. Two of the pairs of variables are conditionally independent. For example,

X is independent of Z, controlling for Y;
Y is independent of Z, controlling for X;
X and Y are associated, controlling for Z.

3. One of the pairs of variables is conditionally independent. For example,
   X is independent of Z, controlling for Y;
   X and Y are associated, controlling for Z;
   Y and Z are associated, controlling for X.

4. No pair of variables is conditionally independent, but the association between any two variables is the same at each category of the third. We then say there is ***homogeneous association***.

5. All pairs of variables are associated, but the association between each pair varies according to the category of the third variable. (Recall Section 10.3.)

Each model has a symbol that indicates the pairs of variables that are associated. Associated variables appear together in the symbol. For instance, (XY, Z) denotes the model for case 2 (above) in which X and Y are associated but the other two pairs are conditionally independent. The symbol (XY, XZ, YZ) denotes the model for case 4, in which all three pairs are associated. Table 15.12 lists the symbols for the models described above. All the models provide some structure for the pattern of association except for the one symbolized by (XYZ). This model fits any sample three-way table perfectly, allowing the associations to be nonhomogeneous. It is called the ***saturated model***.

Table 15.12: Some Loglinear Models for Three-Dimensional Contingency Tables

Model Symbol	Interpretation
(X, Y, Z)	All pairs are conditionally independent.
(XY, Z)	X and Y the only associated pair.
(XY, YZ)	X and Z the only conditionally independent pair.
(XY, YZ, XZ)	Each pair associated, controlling for the third variable. Homogeneous association.
(XYZ)	All pairs associated, but nonhomogeneous.

Odds Ratio Interpretations for Loglinear Models

Interpretations of associations in loglinear models, like those in logistic regression models, can use the odds ratio. In 2×2 contingency tables, independence is equivalent to a population odds ratio of 1.0. In a three-way table, ***conditional independence*** between X and Y means that the population odds ratios in the Xy partial tables all equal 1.0. ***Homogeneous association*** means that the population odds ratios in the XY partial tables are identical at each category of Z.

For instance, a $2 \times 2 \times 3$ table consists of three partial tables each of size 2×2, with two categories for X and two categories for Y measured at three levels of Z. Suppose the XY population odds ratio is 1.0 at the first level of Z, 1.0 at the second level of Z, and 1.0 at the third level of Z. Then, X and Y are conditionally independent, and loglinear model (XZ, YZ) holds. If the population odds ratio = 2.2 at the first level of Z, 2.2 at the second level of Z, and 2.2 at the third level of Z, then there is conditional association but it is homogeneous. So, model (XY, XZ, YZ) holds. When the XY odds ratios are the same at all levels of Z, necessarily the XZ odds ratios are the same at all levels of Y and the YZ odds ratios are the same at all levels of X.

In fitting loglinear models, software provides expected frequencies having odds ratios that perfectly satisfy the model. If the model fits well, these odds ratios help us interpret the associations implied by the model.

Example 15.9 Estimated Odds Ratios for Substance Use Data

Denote the variables in Table 15.11 by A for alcohol use, C for cigarette use, and M for marijuana use. Table 15.13 contains expected frequencies for the loglinear model (AC, AM, CM). This model permits an association for each pair of variables. But, it assumes homogeneous association, in the sense that the odds ratio between two variables is the same at each level of the third variable. The expected frequencies are very close to the observed counts, so the model seems to fit well.

Let's study the estimated association between cigarette use and marijuana use, controlling for alcohol use, using the expected frequencies. For those who used alcohol, the estimated odds ratio between C and M equals

$$\frac{910.4 \times 455.4}{538.6 \times 44.6} = 17.3.$$

Similarly, for those who had not used alcohol, the estimated odds ratio between C and M equals

$$\frac{3.6 \times 279.6}{42.4 \times 1.4} = 17.3.$$

Table 15.13: Expected Frequencies for Loglinear Model (AC, AM, CM) for Alcohol (A), Cigarette (C), and Marijuana (M) Use for High School Seniors

Alcohol Use	Cigarette Use	Marijuana Use Yes	Marijuana Use No
Yes	Yes	910.4	538.6
	No	44.6	455.4
No	Yes	3.6	42.4
	No	1.4	279.6

For each category of A, students who have smoked cigarettes have estimated odds of having smoked marijuana that are 17.3 times the estimated odds for students who have not smoked cigarettes. The model assumes homogeneous association, so the estimated odds ratio is the same at each category of A.

Software for loglinear models provides tables of model parameter estimates from which one can obtain estimated odds ratio. Table 15.14 illustrates. (Software also provides estimates for intercept and main effect parameters. They are not shown here, because interpretations use the interaction terms to describe conditional associations.) For each pair of variables, the association parameter estimate refers to the *log* odds ratio. For the CM conditional association, therefore, the estimated odds ratio at each level of A equals $e^{2.848} = 17.3$. Similarly, the estimated odds ratio equals $e^{2.054} = 7.8$ between A and C at each level of M, and the estimated odds ratio equals $e^{2.986} = 19.8$ between A and M at each level of C. The estimated conditional association is very strong between each pair of variables.

□

Table 15.14: Printout of Association Parameter (Log Odds Ratio) Estimates for Loglinear Model (AC, AM, CM) for Substance Use Data

```
Parameter  Estimate   Std. Error     z      Sig.
  A*C       2.0545      0.1741    11.80   .000
  A*M       2.9860      0.4647     6.43   .000
  C*M       2.8479      0.1638    17.38   .000
```

The model (AC, AM, CM) permits conditional association for each pair of variables. Other possible loglinear models for these data delete at least one of the associations. To illustrate the association patterns implied by some of these models, Table 15.15 presents estimated conditional odds ratios for the expected frequencies for the models. For example, the entry 1.0 for the AC conditional association for model (AM, CM) is the common value of the estimated AC odds ratios at the two categories of M. This model implies conditional independence between alcohol use and cigarette use, controlling for marijuana use. It has estimated odds ratios of 1.0 for the AC conditional association.

Table 15.15: Summary of Estimated Conditional Odds Ratios for Various Loglinear Models Fitted to Substance Use Data

Model		Conditional Odds Ratio		
		AC	AM	CM
(A, C, M)		1.0	1.0	1.0
(AC, M)		17.7	1.0	1.0
(AM, CM)		1.0	61.9	25.1
(AC, AM, CM)		7.8	19.8	17.3
(ACM)	Level 1	13.8	24.3	17.5
(ACM)	Level 2	7.7	13.5	9.7

Table 15.15 shows that estimated conditional odds ratios equal 1.0 for each pairwise term not appearing in a model, such as the AC association in model (AM, CM). The odds ratios for the sample data are those reported for the saturated model (ACM), which provides a perfect fit. For that model, the odds ratios between two variables are not the same at each level of the third variable, so they are reported separately for each level. In each case, they are strong at both levels.

Table 15.15 shows also that estimated conditional odds ratios can vary dramatically from model to model. This highlights the importance of good model selection. An estimate from this table is informative only to the extent that its model fits well. The next section shows how to check model goodness of fit.

15.7 Model Goodness-of-Fit Tests for Contingency Tables*

A ***goodness-of-fit test*** for a model is a test of the null hypothesis that that model truly holds in the population of interest. Section 8.2 introduced the chi-squared test of independence for contingency tables. That test is a goodness-of-fit test for the loglinear model which states that the two categorical variables are statistically independent. The chi-squared statistic compares the observed frequencies to the expected frequencies that satisfy the independence model. Likewise, logistic regression and loglinear models for multidimensional contingency tables have chi-squared goodness-of-fit tests.

Chi-Squared Goodness-of-Fit Statistics

Each model for a contingency table has a set of cell expected frequencies, which are numbers that perfectly satisfy the model and give the best fit to the observed counts. The model goodness of fit is tested by comparing the expected frequencies, denoted by $\{f_e\}$, to the observed frequencies $\{f_o\}$. Chi-squared test statistics summarize the discrepancies. Larger differences between the $\{f_o\}$ and $\{f_e\}$ yield larger values of the test statistics and stronger evidence that the model is inadequate.

Two chi-squared statistics, having similar properties, are commonly used to do this. The ***Pearson statistic***

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

was introduced in Section 8.2 for testing independence. Another statistic, the ***likelihood-ratio statistic***, is

$$G^2 = 2 \sum f_o \log\left(\frac{f_o}{f_e}\right).$$

It equals the difference between the $(-2\log \ell)$ values for the model being tested and for the most complex model possible. Software for generalized linear models calls this statistic the ***deviance***. Both χ^2 and G^2 statistics equal 0 when there is a perfect fit (i.e., all $f_o = f_e$). Since large values indicate a poor fit, the P-value for testing a model is the right-tail probability above the observed value.

If the model truly holds, both test statistics have approximate chi-squared distributions. The degrees of freedom (*df*) for the statistics depend on the model fitted. The *df* resemble the error *df* in regression, equaling the number of responses modeled on the left-hand side of the equation minus the number of parameters on the right-hand side of the model. For logistic regression models, for instance, the number of responses modeled is the number of sample logits for the model. This equals the number of combinations of levels of explanatory variables having observations on the binary response, since there is one logit for each combination. Thus, df = number of logits modeled − number of parameters. The simpler the model, in the sense of fewer parameters, the larger the df for the test.

The chi-squared approximation is better for larger sample sizes. The Pearson statistic is preferred when the expected frequencies average between about 1 and 10, but neither statistic works well if most of the expected frequencies are less than about 5. It is not appropriate to use these tests if any of the predictors are not categorical. The chi-squared sampling distributions result only when they are applied to contingency tables with relatively large counts.

Example 15.10 Logistic Model Goodness of Fit for Death Penalty Data

Examples 15.2 (page 694) and 15.5 (page 702) used the model

$$\text{logit}[P(y = 1)] = \alpha + \beta_1 d + \beta_2 v$$

to describe how the probability of the death penalty depends on defendant's race d and victims' race v. A goodness-of-fit test analyzes whether this model with main effects is adequate for describing the data (Table 15.3). The more complex model containing the interaction term is necessary if the main effects model fits poorly.

Software automatically finds the model expected frequencies and the goodness-of-fit statistics. For instance, this model estimated a probability of 0.233 that a black defendant receives the death penalty for having white victims. Table 15.3 showed there were 48 such black defendants, so the expected number receiving the death penalty equals $48(0.233) = 11.2$. This is the expected frequency for the cell in the table having observed frequency 11.

The df equal the number of logits minus the number of parameters in the model. The death penalty data have four logits, one for each combination of defendant's race and victims' race. The model has three parameters, so both goodness-of-fit statistics have

$$df = \text{Number of logits} - \text{number of parameters} = 4 - 3 = 1.$$

Table 15.16 shows the results of chi-squared goodness-of-fit tests for the logistic model for the death penalty data. The null hypothesis for the tests is that the logistic model with main effects truly holds; that is, no interaction occurs between defendant's race and victims' race in their effects on the death penalty verdict. The Pearson test statistic is $\chi^2 = 0.20$ and the likelihood-ratio test statistic is $G^2 = 0.38$. These test statistic values are small, so neither P-value is small. The model fits the data well. The null hypothesis that the model holds is plausible.

□

Table 15.16: Goodness-of-Fit Tests for Logistic Model with Main Effects Fitted to Death Penalty Data

Goodness-of-fit Tests

	Value	df	Sig.
Likelihood Ratio	.380	1	.538
Pearson Chi-Square	.198	1	.656

Standardized Residuals

The chi-squared statistics provide global measures of lack of fit. When the fit is poor, a closer look at the cells of the table may reveal the nature of the lack of fit. Most software for logistic and loglinear models reports ***standardized residuals*** (sometimes called *adjusted residuals*, which make a cell-by-cell comparison of f_o and f_e. Each standardized residual has form

$$\text{Standardized residual} = \frac{f_o - f_e}{\text{Standard error of } (f_o - f_e)}.$$

When the model truly holds, standardized residuals behave like standard normal variables. A large standardized residual (say, exceeding 3 in absolute value) provides strong evidence of lack of fit in that cell. The standardized residuals presented in Section 8.3 are special cases for the bivariate model of independence.

For the logistic model for the death penalty data, the standardized residuals all equal ± 0.44. They are small and provide no evidence of lack of fit. This is not surprising, since the goodness-of-fit statistics are small. In fact, when $df = 1$ for the goodness-of-fit test, only one standardized residual is nonredundant, and the square of any of them equals the χ^2 test statistic.

Loglinear Model Goodness of Fit

The same goodness-of-fit formulas apply to loglinear models. Likewise, standardized residuals compare individual cell counts to expected frequencies satisfying the model.

Examples 15.8 (page 713) and 15.9 (page 716) used loglinear models to describe associations among alcohol use, cigarette use, and marijuana use, for high school students. Table 15.17 displays results of Pearson χ^2 and likelihood-ratio G^2 goodness-of-fit tests for various loglinear models, ranging from the model (A, C, M) for which each pair of variables is independent to the model (AC, AM, CM) for which each pair is associated but the association between two variables is the same at each level of the third. The smaller the chi-squared statistics, the better the fit. Small P-values contradict the null hypothesis that the model is adequate. It is usually preferable to select the simplest model that provides a decent fit to the data. If no model fits well, the standardized residuals highlight cells contributing to the lack of fit.

Table 15.17: Goodness-of-Fit Tests for Loglinear Models of Alcohol (A), Cigarette (C), and Marijuana (M) Use, with Likelihood-ratio (G^2) and Pearson (χ^2) Chi-squared Test Statistics

Model	G^2	χ^2	*df*	P-Value
(A, C, M)	1286.0	1411.4	4	¡ 0.0001
(A, CM)	534.2	505.6	3	.000
(C, AM)	939.6	824.2	3	.000
(M, AC)	843.8	704.9	3	.000
(AC, AM)	497.4	443.8	2	.000
(AC, CM)	92.0	80.8	2	.000
(AM, CM)	187.8	177.6	2	.000
(AC, AM, CM)	0.4	0.4	1	0.54

The chi-squared distribution falls on the positive part of the line with mean equal to the *df* value. Hence, a value such as $G^2 = 1286.0$ based on $df = 4$ falls way out in the right-hand tail. It has a tiny P-value and indicates a very poor fit. From Table 15.17, the only model that passes the goodness-of-fit test is (AC, AM, CM). This model allows association between all pairs of variables but assumes that the odds ratio between each pair is the same at each category of the third variable. The models that lack any associations fit poorly, having P-values of 0.000.

Comparing Models by Comparing G^2 Values

Table 15.17 illustrates two important properties of the likelihood-ratio G^2 statistic. First, G^2 has similar properties as the SSE (sum of squared residuals) measure in regression. Both compare observed responses to values expected if a model holds, and both cannot increase as the model becomes more complex. For instance, (A, CM) is a more complex model than (A, C, M), since it allows one association. Hence, it provides a better fit and its G^2-value is smaller. Similarly, G^2 drops further for model (AC, CM) and further yet for (AC, AM, CM). The Pearson χ^2 statistic, unlike the likelihood-ratio G^2, does not have this property. It could potentially increase as a model gets more complex, although in practice this rarely happens.

The second property of G^2 refers to model comparison. This was first discussed at the end of Section 15.3. Section 11.6 introduced an F test comparing complete and reduced regression models, based on the reduction in SSE. A similar test comparing models for categorical responses uses the reduction in G^2 values. To test the null hypothesis that a model truly holds versus the alternative hypothesis that a more complex model fits better, the test statistic is the difference in G^2 values. This difference is identical to the difference in $(-2\log \ell)$ values for the two models. It is a chi-squared statistic with degrees of freedom equal to the difference in *df* values for the two G^2 values. This is the likelihood-ratio test for comparing the models.

To illustrate, we compare models (AC, CM) and (AC, AM, CM). We test the null hypothesis that the reduced model (AC, CM) is adequate against the alternative that the more complex model (AC, AM, CM) is better. The likelihood-ratio test analyzes whether we can drop the AM association from model (AC, AM, CM). The test statistic is the difference between their G^2 values, $92.0 - 0.4 = 91.6$, based on $df = 2 - 1 = 1$. This chi-squared statistic has a P-value of $P = 0.000$. So, model (AC, AM, CM) fits significantly better than (AC, CM).

It is not possible to use G^2 to compare a pair of models such as (A, CM) and (AC, AM). Neither is a special case of the other, since each allows association that the other excludes.

Connection Between Logistic and Loglinear Models

Any logistic regression model for a contingency table has an equivalent loglinear model. That loglinear model contains the same associations as the logistic model does between the response variable and the explanatory variables, and it contains the most general term for describing relationships among the explanatory variables.

Example 15.11 Loglinear Models for Death Penalty Data

Examples 15.2 (page 694) and 15.5 (page 702) used a logistic model for the death penalty data having main effects for defendant's race and victims' race. It provides a good fit, with $G^2 = 0.38$ and $\chi^2 = 0.20$ based on $df = 1$.

Let $D =$ defendant's race, $V =$ victims' race, and $P =$ death penalty verdict. For that logistic model, P is the response variable, and D and V have effects on P but without interaction. It has precisely the same fit as the loglinear model that also allows an association between D and P and between V and P and also has a term to allow an association between the explanatory variables D and V. That is the loglinear model

(DP, VP, DV) that allows associations for each pair of variables. That loglinear model has the same values for G^2 and χ^2 for testing fit, the same expected frequencies, and the same estimated odds ratios between each predictor and the death penalty outcome. You can check this, as an exercise, by using software to fit this loglinear model to the data (Table 15.3).

□

Distinction Between Logistic and Loglinear Models

Logistic models distinguish between a single response variable and a set of explanatory variables. By contrast, loglinear models treat every variable as a response variable. Although designed for a different purpose, logistic models imply association patterns among variables that are equivalent to ones expressed by loglinear models. For data in contingency tables, similar conclusions result with the two approaches.

Most applications have a single response variable. It is then more natural to use a logistic regression model than a loglinear model. The logistic analysis focuses on the effects of the predictors on the response, much as in ordinary regression modeling. For that reason, logistic regression is more important in practice. That is also why we gave it more attention than loglinear models in this chapter.

15.8 Chapter Summary

Chapter 8 presented methods for analyzing association between two categorical variables. The methods of this chapter showed how to model a categorical response variable in terms of possibly *several* explanatory variables.

- For binary response variables, the ***logistic regression*** model describes how the probability of a particular category depends on explanatory variables. It uses a linear model for the ***logit*** transform of the probability, which is the log of the odds. For a quantitative explanatory variable, an S-shaped curve describes how the probability changes as the predictor changes.

- The antilog of a $\hat{\beta}$ parameter estimate in logistic regression is a multiplicative effect on the odds for the response variable, for each 1-unit increase in the predictor variable of which it is a coefficient. Thus, for logistic regression the ***odds ratio*** is a natural measure of the nature and strength of an association.

- A parameter value of $\beta = 0$ corresponds to a predictor having no effect on the response. To test H_0: $\beta = 0$, the ***Wald test*** uses the squared ratio of $\hat{\beta}$ to its standard error. The ***likelihood-ratio test*** compares values of $(-2 \log \ell)$ for models with and without that term, where ℓ is the maximized *likelihood* function. The large-sample distribution of these test statistics is chi-squared with *df* = 1.

- For *ordinal* response variables, an extension of logistic regression uses *cumulative logits*, which are logits of *cumulative probabilities*. The model is called a ***cumulative logit model***. The effects of predictors are the same for each cumulative probability.

- For *nominal* response variables, an extension of logistic regression forms logits by pairing each category with a baseline category. Each logit equation has separate parameters.

- ***Loglinear*** models are useful for investigating association patterns among a set of categorical response variables. They consider possible conditional independence patterns and use conditional odds ratios to describe association.

- For models for contingency tables, Pearson and likelihood-ratio chi-squared statistics test the goodness of fit of the model to the data.

Karl Pearson introduced the chi-squared test for bivariate contingency tables in 1900. The models presented in this chapter did not become popular until near the end of the 1900s. They are examples of ***generalized linear models***, which apply to discrete as well as continuous response variables. The statistician and sociologist Leo Goodman is responsible for many of the developments in this area. Social scientists now have available a wide variety of tools for analyzing categorical data.

PROBLEMS

Practicing the Basics

1. A logistic regression model describes how the probability of voting for the Republican candidate in a presidential election depends on x, the voter's total family income (in thousands of dollars) in the previous year. The prediction equation for a particular sample is

$$\log\left[\frac{\hat{P}(y=1)}{1-\hat{P}(y=1)}\right] = -1.00 + 0.02x.$$

a) Identify $\hat{\beta}$ and interpret its sign.
b) Find the estimated probability of voting for the Republican candidate when (i) income = 10 thousand, (ii) income = 100 thousand.
c) At which income level is the estimated probability of voting for the Republican candidate (i) equal to 0.50? (ii) greater than 0.50?
d) For the region of x values for which $P(y = 1)$ is near 0.50, give a linear approximation for the change in the probability for an increase of one thousand dollars in income.
e) Explain the effect of a thousand dollar increase in family income on the odds of voting Republican.

2. Refer to the previous exercise. When the explanatory variables are x_1 = family income, x_2 = number of years of education, and s = sex (1 = male, 0 = female), the prediction equation is

$$\text{logit}[\hat{P}(y = 1)] = -2.40 + 0.02x_1 + 0.08x_2 + 0.20s.$$

For this sample, x_1 ranges from 6 to 157 with a standard deviation of 25, and x_2 ranges from 7 to 20 with a standard deviation of 3.
a) Find the estimated probability of voting Republican for (i) a man with 16 years of education and income 30 thousand dollars, (ii) a woman with 16 years of education and income 30 thousand dollars.
b) Convert the probabilities in (a) to odds, and find the odds ratio, the odds for men divided by the odds for females. Interpret.
c) Show how the odds ratio in (b) relates to the sex effect in the prediction equation.
d) Holding the other variables constant, find the estimated effect on the odds of voting Republican of
i) A standard deviation change in x_2.
ii) A standard deviation change in x_1.
Which predictor has the larger standardized effect? Interpret.

3. A sample of 54 elderly men are given a psychiatric examination to determine whether symptoms of senility are present. A subtest of the Wechsler Adult Intelligence Scale (WAIS) is the explanatory variable. The WAIS scores range from 4 to 20, with a mean of 11.6. Higher values indicate more effective intellectual functioning. Table 15.18 shows results.
a) Report the prediction equation. Explain why it suggests that the probability of senility decreases at higher levels of the WAIS.
b) Show (i) $\hat{P}(y = 1) = 0.50$ at $x = 7.2$, (ii) $\hat{P}(y = 1) < 0.50$ for $x > 7.2$.
c) Estimate the probability of senility at $x = 20$.
d) The fit of the linear probability model is $\hat{P}(y = 1) = 0.847 -$

$0.051x$. Estimate the probability of senility at $x = 20$. Does this make sense?
e) Test H_0: $\beta = 0$ against $H_a : \beta \neq 0$. Report and interpret the P-value.

Table 15.18:

Variable	B	Std. Error	Wald Chi-square	Sig.
INTERCEPT	2.0429	1.0717	3.6338	0.0566
WAIS	-0.2821	0.1007	7.8487	0.0051

4. Refer to the previous exercise. One of the WAIS subtests, called *Picture completion*, asks questions about 20 pictures that have one vital detail missing. It is considered a test of attention to fine detail. The observations for 20 subjects on (x, y), where $x =$ picture completion score and $y =$ symptoms of senility (1 = yes) are:

   (7, 1), (5, 1), (3, 1), (8, 1) (1, 1), (2, 1) (9, 1), (3, 1),
   (6, 1), (4, 1), (6, 0), (9, 0), (7, 0), (7, 0), (10, 0), (12, 0),
   (14, 0), (8, 0), (8, 0), (11, 0)

   a) Find the prediction equation for logistic regression.
   b) Find the estimated probability that symptoms of senility are present when (i) $x = 0$, (ii) $x = 20$.
   c) Over what range of x-scores is the estimated probability of senility greater than 0.50?
   d) Estimate the effect of a one-unit increase in x on the odds that senility symptoms exist.

5. The final subsection of Section 15.2 used estimated probabilities to describe the effect of husband's earnings on the decision to buy a home. Perform a similar analysis to describe the effect of wife's earnings, when husband's earnings = \$50,000, years married = 3, the wife is working in two years, number of children = 0, add child in two years = 0, head's education = 16 years, and parents' home ownership = 0. Interpret.

6. Table 12.1 in Chapter 12 reported GSS data on political ideology by party affiliation of

	1	2	3	4	5	6	7
Democrat	9	20	17	36	4	5	0

Republican	0	2	7	23	23	17	2

Use logistic regression to describe the effect of ideology on the probability of being a Democrat.
a) Report the prediction equation, and estimate the probability of Democratic affiliation at ideology level (i) 1 = extremely liberal, (ii) 7 = extremely conservative.
b) Use the model to test whether the variables are independent. Report the test statistic, P-value, and interpret.
c) Use the odds ratio to describe the effect on party affiliation of a change in ideology from (i) 1 = extremely liberal to 2 = liberal, (ii) 1 = extremely liberal to 7 = extremely conservative.
d) Construct and interpret a 95% confidence interval for the population odds ratio in (c), case (i).

7. A study of mother's occupational status and children's schooling[3] reported the prediction equation

$$\text{logit}[\hat{P}(y=1)] = 0.75 + 0.35b + 0.13f + 0.09m + 0.30fo + 0.21mo - 0.92me - 0.16s,$$

where $P(y=1)$ is the probability the child obtains a high school degree, b = respondent's year of birth, f = father's education, m = mother's education (0 to 17), fo = father's occupational level, mo = mother's occupational level (1 to 9), me = whether mother employed (1 = yes), s = number of siblings. All effects were significant at the 0.01 level.
a) Interpret the coefficient of mother's education.
b) Interpret the coefficient of whether mother employed.
c) The author reported that a one-point increase in mother's occupational level is associated with a 24% increase in the odds of a high school diploma. Explain how he made this interpretation.

8. Let $P(y=1)$ denote the probability that a randomly selected respondent supports current laws legalizing abortion, estimated using sex of respondent ($s=0$, male; $s=1$, female), religious affiliation ($r_1=1$, Protestant, 0 otherwise; $r_2=1$, Catholic, 0 otherwise; $r_1=r_2=0$, Jewish), and political party affiliation ($p_1=1$, Democrat, 0 otherwise; $p_2=1$, Republican, 0 otherwise, $p_1=p_2=0$, Independent). The logistic model with main effects has prediction equation

$$\text{logit}[\hat{P}(y=1)] = 0.11 + 0.16s - 0.57r_1 - 0.66r_2 + 0.47p_1.$$

a) Give the effect of sex on the odds of supporting legalized abortion; that is, if the odds of support for females equal θ times the

[3] M. Kalmin, *Amer. Sociolog. Rev.*, Vol. 59, 1994, p. 257

odds of support for males, report $\hat{\theta}$.
b) Give the effect of being Democrat instead of Independent on the estimated odds of support for legalized abortion.
c) Give the effect of being Democrat instead of Republican on the estimated odds of support for legalized abortion.
d) Find the estimated probability of supporting legalized abortion, for (i) female Jewish Democrats, (ii) male Catholic Republicans.

9. Table 15.19 shows results of a study on the effects of AZT in slowing the development of AIDS symptoms. In the study, 338 veterans whose immune systems were beginning to falter after infection with the AIDS virus were randomly assigned either to receive AZT immediately or to wait until their T cells showed severe immune weakness. The response is whether they developed AIDS symptoms during the three-year study. Software reports the fit of the logistic model with main effects:

```
Parameter      Estimate   Std. Error  Wald Chi-Square   Sig
 Intercept      -1.0736     0.2629      16.6705        .0001
 [azt=yes]      -0.7195     0.2790       6.6507        .0099
 [race=black]    0.0555     0.2886       0.0370        .8476
```

a) Set up dummy variables, and report the prediction equation.
b) Interpret the signs of the azt and race estimates.
c) For black veterans without immediate AZT use, estimate the probability of AIDS symptoms.
d) Find the estimated conditional odds ratio between AZT use and the development of symptoms. Interpret.
e) Test for the effect of AZT use. Interpret.

Table 15.19:

		Symptoms	
Race	AZT Use	Yes	No
White	Yes	14	93
	No	32	81
Black	Yes	11	52
	No	12	43

Source: *The New York Times*, February 15, 1991.

10. For Table 15.11, Table 15.20 shows a SAS printout for a logistic model treating marijuana use as the response variable.
a) Set up dummy variables and report the prediction equation.

b) Why are the estimates in the table equal to 0 at the second category of each predictor?
c) Interpret the signs of the effects of alcohol use and cigarette use.
d) Estimate the probability of having used marijuana (i) for those who have not used alcohol or cigarettes, (ii) for those who have used both alcohol and cigarettes.
e) Show how to convert the coefficient for alcohol use to an estimated odds ratio. Interpret.

Table 15.20:

Parameter		DF	Estimate	Std Err	ChiSquare	Pr > Chi
INTERCEPT		1	-5.309	0.4752	124.820	0.0001
ALCOHOL	yes	1	2.986	0.4647	41.293	0.0001
ALCOHOL	no	0	0.000	0.0000	0.	0.
CIGARETT	yes	1	2.848	0.1638	302.141	0.0001
CIGARETT	no	0	0.000	0.0000	0.	0.

11. A sample of inmates being admitted to the Rhode Island Department of Corrections were asked whether they ever injected drugs and were tested for hepatitis C virus (HCV). The number who reported injecting drugs were 306 of the 887 men who tested HCV positive, 61 of the 3044 men who tested HCV negative, 110 of the 197 women who tested HCV positive, and 13 of the 288 women who tested HCV negative. The authors[4] concluded that the prevalence of HCV may be underestimated by testing only those who reported injecting drugs.
a) Report the results as a contingency table.
b) Find and interpret the sample odds ratios between HCV status and whether injected drugs, for each sex.
c) Define dummy variables and specify a model for which the two odds ratios estimated in (b) are identical in the population.
d) Fit the model in (c), and report the model-based odds ratio estimate.

12. Table 15.21 refers to individuals who applied for admission into graduate school at the University of California in Berkeley. Data[5] are presented for five of the six largest graduate departments at the university. The variables are:
A: Whether admitted (yes, no)

[4] G. Macolino et al., *Amer. J. Public Health*, vol. 95, pp. 1739-1740, 2005
[5] from D. Freedman, R. Pisani, and R. Purves, *Statistics*, W. W. Norton, 1978, p. 14

S: Sex of applicant (male, female)
D: Department to which application was sent $(D_1, D_2, D_3, D_4, D_5)$
a) Construct the two-way table for sex and whether admitted, collapsing the table over department. Find the odds ratio and interpret.
b) Treating A as the response and D and S as categorical predictors, fit the logit model having main effects. Report the prediction equation. Interpret the coefficient of S in this equation by finding the estimated conditional odds ratio between A and S, controlling for D.
c) Contrast the model estimated conditional odds ratio between A and S in (b) with the odds ratio reported in (a). Explain why they differ so much, satisfying Simpson's paradox.

Table 15.21:

		Admitted	
Department	Sex	Yes	No
D_1	M	353	207
	F	17	8
D_2	M	120	205
	F	202	391
D_3	M	138	279
	F	131	244
D_4	M	53	138
	F	94	299
D_5	M	22	351
	F	24	317

13. Refer to Table 8.15 on page 334 in Chapter 8. Treating happiness as the response variable, Table 15.22 shows results of fitting the cumulative logit model, using scores (1, 2, 3) for income, and the chi-squared test of independence of Chapter 8.
a) Why does the table report two intercept estimates?
b) Report and interpret the income effect.
c) Using the model, test the hypothesis of no income effect. Report the test statistic and P-value, and interpret. (Similar results are obtained using an ordinal analysis with the gamma measure of association, for which gamma $= 0.228$, $se = 0.114$, $z = 1.96$, and the P-value is 0.050.)
d) Suppose we instead used the Pearson chi-squared test of independence. Report the test statistic and P-value, and compare results to those in (c). Why are they so different?

Table 15.22:

	Estimate	Std. Error	Wald	df	Sig.
INTERCP1 [happy=1]	-.792	.415	3.63	1	.057
INTERCP2 [happy=2]	1.681	.435	14.94	1	.000
INCOME	.418	.223	3.53	1	.060

Statistic	DF	Value	Sig.
Pearson Chi-square	4	3.816	0.431

14. Using software, replicate the results in Section 15.4 for the example on political ideology and party affiliation.

15. Table 15.23 refers to passengers in autos and light trucks involved in accidents in the state of Maine in 1991. The table classifies subjects by sex, location of accident, seat-belt use, and a response variable having categories (1) not injured, (2) injured but not transported by emergency medical services, (3) injured and transported by emergency medical services but not hospitalized, (4) injured and hospitalized but did not die, (5) injured and died.
a) Fit a cumulative logit model having main effects for sex ($s = 1$ for males and 0 for females), location ($r = 1$ for rural and 0 for urban), and seat-belt use ($s = 1$ for yes and 0 for no). State the prediction equation. Interpret the sign of the effect for each predictor.
b) Report and interpret an odds ratio describing the effect of wearing a seat belt.
c) Construct a 95% confidence interval for the true odds ratio for the effect of wearing a seat belt. Interpret.
d) Conduct a test of the hypothesis of no effect of seat belt use on the response, controlling for sex and location. Report the P-value and interpret.

16. For the previous exercise, fit the model that also has the three two-way interactions. Use a likelihood-ratio test to compare this model to the main effects model. Interpret. (*Hint*: The test statistic is the difference between the $(-2\log\ell)$ values.)

17. A baseline-category logit model fit predicting preference for President (Democrat, Republican, Independent) using x = annual income (in \$10,000 dollars) is $\log(\hat{\pi}_D/\hat{\pi}_I) = 3.3 - 0.2x$ and $\log(\hat{\pi}_R/\hat{\pi}_I) =$

Table 15.23:

Sex	Location	Seat-Belt	Response 1	2	3	4	5
Female	Urban	No	7287	175	720	91	10
		Yes	11,587	126	577	48	8
	Rural	No	3246	73	710	159	31
		Yes	6134	94	564	82	17
Male	Urban	No	10,381	136	566	96	14
		Yes	10,969	83	259	37	1
	Rural	No	6123	141	710	188	45
		Yes	6693	74	353	74	12

Source: Dr. Cristanna Cook, Medical Care Development, Augusta, Maine.

$1.0 + 0.3x$.
a) Use an estimated odds ratio to describe how the choice between Republican and Independent depends on income.
b) Explain how to get the prediction equation for $\log(\hat{\pi}_R/\hat{\pi}_D)$. Interpret its slope.

18. For a sample of people in Ithaca, New York, consider the most recent time each person shopped for clothes. You plan to model the choice to shop downtown, at the local mall, or on the Internet, as a function of several variables, such as annual income, whether a student, and distance of residence from downtown. Explain the type of model you would use, and why.

19. Refer to the 3×7 table in Table 12.1 (page 526) on party identification and political ideology.
a) Fit a baseline-category logit model, treating party affiliation as the response. Interpret the political ideology effect for each pair of party IDs.
b) Fit a cumulative logit model, treating political ideology as the response. Interpret the party ID effects.

20. Using software, replicate the results in Section 15.5 for the example on belief in life after death, sex, and race.

21. Consider the fit of the model (AC, AM, CM) to Table 15.11 for the survey of high school seniors.
a) Use the expected frequencies in Table 15.13 to estimate the conditional odds ratios between A and M at each level of C.
b) Show how to obtain the estimated odds ratio in (a) from the parameter estimates for the model in Table 15.14.

c) By contrast, what is the estimated conditional odds ratio between A and M for the loglinear model denoted by (AC, CM)?

22. Refer to the loglinear model analyses reported in Section 15.6 for use of marijuana, alcohol, and cigarettes. Use software to replicate all the analyses shown there.

23. Consider a four-way cross-classification of variables W, X, Y, and Z. State the symbol for the loglinear model in which: **a)** All pairs of variables are independent.
b) X and Y are associated but other pairs of variables are independent.
c) All pairs of variables are associated, but the conditional associations are homogeneous.

24. For the logistic model reported in Exercise 9 above on AZT use and AIDS, software reports a Pearson goodness-of-fit statistic equal to 1.39 with $df = 1$, with P-value 0.24. Specify H_0 for this test, and interpret the P-value.

25. Refer to the survey data for high school seniors in Table 15.11 (page 714).
a) Specify a logistic regression model that would have fit exactly the same as the loglinear model (AC, AM, CM).
b) The Pearson and likelihood-ratio goodness-of-fit statistics for the model in (a) both equal 0.4, with $df = 1$. Interpret.

Concepts and Applications

26. Refer to the "Student survey" data file (Exercise 1.11). Using software, conduct and interpret a logistic regression analysis using $y =$ opinion about abortion with predictors
a) political ideology.
b) sex and political ideology.

27. Refer to the data file you created in Exercise 1.12. For variables chosen by your instructor, fit a logistic regression model and conduct descriptive and inferential statistical analyses. Interpret and summarize your findings.

28. The data given in Exercise 14 in Chapter 10 came from an early study on the death penalty and racial characteristics. Analyze those data using methods of this chapter. Summarize your main findings in a way that you could present to the general public, using as little technical jargon as possible.

29. According to the *Independent* newspaper (London, March 8, 1994), the Metropolitan Police in London reported 30,475 people as missing in the year ending March 1993. For those of age 13 or less, 33 of 3271 missing males and 38 of 2486 missing females were still missing a year later. For ages 14–18, the values were 63 of 7256 males and 108 of 8877 females; for ages 19 and above, the values were 157 of 5065 males and 159 of 3520 females. Analyze and interpret these data. (Thanks to Dr. P.M.E. Altham for showing me these data.)

30. In a study of whether an educational program makes sexually active adolescents more likely to obtain condoms, adolescents were randomly assigned to two experimental groups. The educational program, involving a lecture and videotape about transmission of the HIV virus, was provided to one group but not the other. In logistic regression models, factors observed to influence a teenager to obtain condoms were sex, socioeconomic status, lifetime number of partners, and the experimental group. Table 15.24 summarizes results.
a) Find the parameter estimates for the fitted model, using (1, 0) dummy variables for the first three predictors.
b) Explain why either the estimate of 1.38 for the odds ratio for sex or the corresponding confidence interval seems incorrect. The confidence interval is based on taking antilogs of endpoints of a confidence interval for the log odds ratio. Show that if the reported confidence interval is correct, then 1.38 is actually the *log* odds ratio, and the estimated odds ratio equals 3.98.

Table 15.24:

Variables	Odds Ratio	95% Confidence Interval
Group (education versus none)	4.04	(1.17, 13.9)
Sex (males versus females)	1.38	(1.23, 12.88)
SES (high versus low)	5.82	(1.87, 18.28)
Lifetime number of Partners	3.22	(1.08, 11.31)

Source: V. I. Rickert et al., *Clinical Pediatrics*, Vol. 31 (1992), pp. 205–210.

31. A Canadian survey of factors associated with whether a person is a hunter of wildlife showed the results in Table 15.25. Explain how to interpret the results in this table. The study abstract[6] stated, "Men are 10 times more likely to hunt wildlife than females." Comment on how this conclusion was reached, and whether it is correct.

[6]by R. Mitchell, *Crossing Boundaries*, vol. 1, 107-117, 2001

Table 15.25:

	B	S.E.	Wald	Sig	Exp(B)
Constant	-5.04	0.16	943.1	.000	
Male	2.34	0.15	259.9	.000	10.39
Live in rural area	0.98	0.10	106.2	.000	2.67
Not married	-0.04	0.12	0.1	.717	.96
Not employed	-0.36	0.12	8.8	.003	.70
Age: 15 to 29	0.21	0.13	2.5	.113	1.24
Age: 50 or more	-0.27	0.12	4.7	.030	.77
Education up to HS	0.38	0.10	14.9	.000	1.46
Naturalist club member	1.64	0.11	228.6	.000	.42

32. A study[7] compared the relative frequency of mental health problems of various types among U.S. Army members before deployment to Iraq, U.S. Army members after serving in Iraq, U.S. Army members after serving in Afghanistan, and U.S. Marines after serving in Iraq. The study stated, "Potential differences in demographic factors among the four study groups were controlled for in our analysis with the use of logistic regression." For this study, identify the response variable, the primary explanatory variable, and likely control variables.

33. A 2006 report by the Office of Applied Studies for the Substance Abuse and Mental Health Services Administration (http://www.oas.samhsa.gov/) about factors that predict marijuana use stated, "Multiple logistic regression also confirmed that the risk of recent marijuana initiation increased with increasing age among youths aged 12 to 14, but the risk decreased with increasing age among those aged 15 to 25." What does this suggest about the way that age appears in the model used?

34. Refer to the data in Exercise 16 (page 348) in Chapter 8 on happiness and marital status. Using a cumulative logit model, analyze and interpret these data.

35. Refer to Table 15.4 on page 694. Show that the association between the defendant's race and the death penalty verdict satisfies Simpson's paradox. What causes this?

36. For a person, let $y = 1$ represent death during the next year and $y = 0$ represent survival. For adults in the UK and in the U.S., the

[7] C. Hoge et al., *New England J. Medic.*, vol. 351, 13-21, 2004

probability of death is well approximated by the model, $\text{logit}[P(y = 1)] = -10.5 + 0.1x$, where x = age in years. Show how the probability of death in the next year increases as x increases from 20 to 60 to 100.

37. State the symbols for the loglinear models for categorical variables that are implied by the causal diagrams in Figure 15.5.

Figure 15.5:

((Fig. 15.5 in 3e))

38. * For the logistic regression model, from the linear approximation $\beta/4$ for the rate of change in the probability at the x-value for which $P(y = 1) = 0.50$, show that $1/|\beta|$ is the approximate distance between the x-values at which $P(y = 1) = 1/4$ (or $P(y = 1) = 3/4$) and at which $P(y = 1) = 1/2$. Thus, the larger the value of $|\beta|$, the less the x-distance over which this change in probability occurs.

39. * For a two-way contingency table, let r_i denote the ith row total, let c_j denote the jth column total, and let n denote the total sample size. Section 8.2 showed that the cell in row i and column j has $f_e = r_i c_j / n$ for the independence model. Show that the log of the expected frequency has an additive formula with terms representing the influence of the ith row total, the jth column total, and the sample size. This formula is the loglinear model for independence in two-way contingency tables.

40. * Explain what is meant by the absence of statistical interaction in modeling the relationship between a response variable y and two explanatory variables x_1 and x_2 in each of the following cases. Use graphs or tables to illustrate.
a) y, x_1, and x_2 are quantitative.
b) y and x_1 are quantitative; x_2 is categorical.
c) y is quantitative; x_1 and x_2 are categorical.
d) y, x_1, and x_2 are binary.
e) y is binary, x_1 and x_2 are quantitative.