15

Logistic Regression Basics

Learning Objectives

In this chapter, you will learn:

- What a logistic regression is
- What odds are and how they are calculated
- What an odds ratio is and how to interpret it
- How to calculate a predicted probability
- How to interpret Nagelkerke's R^2

Introduction

In previous chapters, you learned how to use multiple linear regression to make predictions about a ratio-level dependent variable. In this chapter, you will learn how regression can be used to make predictions about a dichotomous dependent variable. Recall that dichotomous variables have only two values—"1" and "0". Because they have only two values, dichotomous variables are sometimes called binary variables. The type of regression that is used with dichotomous dependent variables is called **logistic regression**, or binary logistic regression. This chapter describes the concepts that underlie logistic regression and explains how to interpret logistic regression results. Since many things that social science researchers want to make predictions about cannot be captured in ratio-level variables, logistic regression is widely used and logistic regression results regularly appear in published reports and articles.

logistic regression A type of regression used to make predictions about a dichotomous dependent variable.

This research focus of this chapter is food security and insecurity. The Universal Declaration of Human Rights enshrines the "right to food" as a basic human right (Article 25). The concept of food security expands on the basic "right to food" by incorporating the idea of consistent access to appropriate food. The United Nations Food and Agriculture Organization (UNFAO) defines food security as "a situation that exists when people [have] secure access to sufficient amounts of safe and nutritious food for normal growth and development and an active, healthy life" (2001). Health Canada measures food security by asking people whether they can afford enough food to eat balanced meals, to maintain their body weight, and to avoid skipping meals or being hungry. (See the "Spotlight on Data" box in this chapter for more information.)

Giannis Papanikos/Shutterstock

Photo 15.1 Some food-insecure households turn to food banks for help. Food bank use in Canada has been at record levels since 2008 (Food Banks Canada 2015).

Like many industrialized countries, Canada has struggled to develop a coherent food policy (MacRae 2012). Over the past several decades, the federal government has begun work on several food-related initiatives, including a national food strategy in 1977–78, an "Action Plan for Food Security" in 1998, and a "National Food Policy Framework" in 2005; however, all were either abandoned or left incomplete (MacRae 2012). Most recently, in 2017, the federal Department of Agriculture and Agri-Food initiated a series of consultations with stakeholders and the public oriented towards developing "A Food Policy for Canada," with the support of several other federal departments and agencies.

The rate of food insecurity in Canada has remained stable since 2007: about 5 per cent of children and 8 per cent of adults live in food-insecure households (Roshanafshar and Hawkins 2015). In Canada, geography is strongly related to food security. For example, the high cost of transportation and storage can make food prices in rural and northern communities prohibitively high and access to fresh foods, difficult. Nunavut has the highest rate of food insecurity in Canada, where more than one in three households (37 per cent) are unable to access to the variety or quantity of food that they need due to lack of money (Roshanafshar and Hawkins 2015). Some food-insecure households rely on food banks to bridge the gap, but they often do not receive enough support to meet their nutritional needs (Tarasuk, Dachner, and Loopstra 2014). Food insecurity can have profound negative effects on people's everyday lives and experiences. As a result, people who are food insecure tend to have poorer physical and mental health than those who are food secure (Vozoris and Tarasuk 2003).

In this chapter, statistical analysis is used to discover the following:

- How are age, gender, personal income, and region of residence related to food insecurity?
- How does the probability of being food insecure change in relation to people's income?
- How does the probability of being food insecure change in relation to people's age?

Spotlight on Data

The Canadian Community Health Survey (Annual Component)

The analyses in this chapter use data from the 2012 Canadian Community Health Survey (CCHS), which was described in Chapter 7. The purpose of the CCHS is to support health monitoring and surveillance programs and to inform health policy at the municipal, provincial, and national levels. Although the CCHS collects data continuously, the questions about food security are sometimes optional and, thus, only asked in some provinces and territories. The 2012 CCHS was the last time that the food security questions were mandatory, and information was collected from people in all provinces and territories. In 2012, the overall response rate for the main component of the CCHS was 67 per cent; overall, 61,707 people completed the survey (Statistics Canada 2013). Only people aged 15 and older are included in the analyses in this chapter; they are considered adults for the purpose of assessing food security.

Statistics Canada relies on the definition of food security established by Health Canada. Households that are food insecure are ones that, in the past year, were "uncertain of having, or unable to acquire, enough food to meet the needs of all their members because they had insufficient money for food." Food-insecure households are divided conceptually into those that are moderately food insecure, that is, they had to compromise in the quality and/or quantity of food consumed, and those that are severely food insecure, that is, they had reduced food intake and disrupted eating patterns (Health Canada 2012).

The food security of adults in Canada is assessed by asking respondents whether any of the following occurred during the past year:

- You and other household members worried food would run out before you got money to buy more.
- The food you and other household members bought just didn't last and there wasn't any money to get more.

Continued

- You and other household members couldn't afford to eat balanced meals.
- You or other adults in your household cut the size of meals or skipped meals because there wasn't enough money to buy food.
- You (personally) ate less than you felt you should because there wasn't enough money to buy food.
- You (personally) were hungry but did not eat because you couldn't afford enough food.
- You (personally) lost weight because you didn't have enough money for food.
- You or other adults in your household did not eat for a whole day because there wasn't enough money to buy food.

Adults who answered that this was "often" or "sometimes" true, or that it occurred in three or more months, for two to five items in this list are considered to be moderately food insecure. Adults who gave these answers for six or more items in this list are considered to be severely food insecure.

Understanding the Conceptual Framework of Logistic Regression

Although linear regression is useful for making predictions about ratio-level dependent variables, social science researchers regularly study things that are captured in dichotomous variables. For instance, researchers might be interested in predicting whether or not people are unemployed, whether or not they have a student loan, or whether or not they voted in the last election.

Theoretically, linear regression can be used to find the straight line that best fits the pattern of the relationship between a ratio-level independent variable and a dichotomous dependent variable. In this situation, the line of best fit might look something like the dotted line in Figure 15.1. As with any other linear regression, the line of best fit is the one that minimizes the sum of the squared distances between each case and the line.

Although it is technically possible to use this approach, the results are problematic for several reasons. First, the line of best fit predicts values on the dependent variable (y) that are above "1" or below "0". If researchers want to predict whether or not people voted in the last election, predicting values above "yes" or values below "no" doesn't make sense. Second, the line of best fit predicts values in between "0" and "1", which aren't legitimate attributes. Because of this, the unstandardized slope coefficient of the independent variable (x) isn't particularly meaningful. For example, the regression line in Figure 15.1 shows that every one-unit increase in the independent variable is associated with a 0.02 increase in the dependent variable.

Statisticians work around both of these problems by transforming dichotomous dependent variables before using them in regressions. In Chapter 14, you learned that when a variable is transformed, the values on the original variable are replaced with values that are a mathematical function of the original value. In

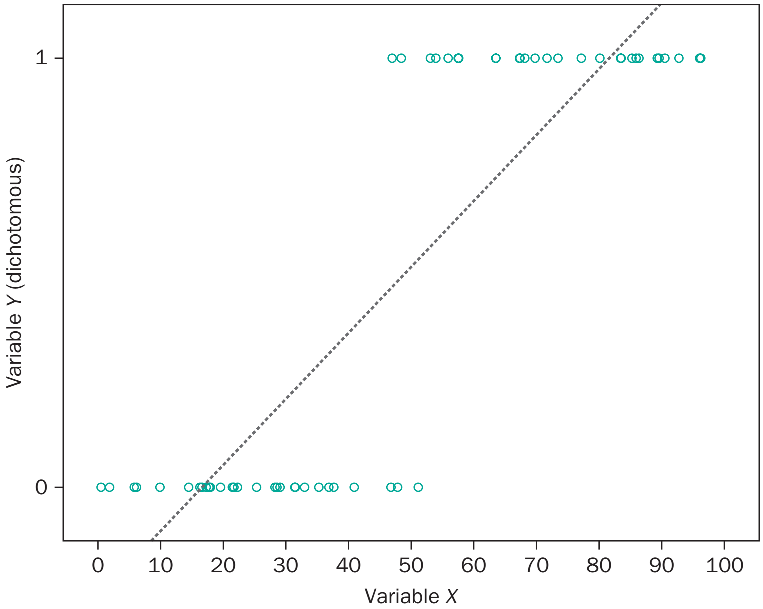


Figure 15.1 Using Linear Regression to Predict the Relationship between a Ratio-Level Independent Variable and a Dichotomous Dependent Variable

this section, you'll learn how dichotomous dependent variables can be transformed mathematically so that they range from negative infinity to positive infinity. This transformation allows researchers to continue using the same general approach as they use in linear regression to make predictions about a dichotomous dependent variable—although the transformation of the dependent variable changes the interpretation of the regression coefficients.

The mathematical transformation of the dichotomous dependent variable in a logistic regression has three steps, which I'll describe in sequence. The first step represents an important conceptual shift: instead of trying to predict the *value* on the dependent variable, researchers try to predict the *probability* that a case will have the value "1" on the dependent variable. In the context of logistic regression, researchers usually construct the dependent variable so that having a "1" value indicates the presence or occurrence of the thing they are interested in (e.g., unemployment, student debt, voting). In the analyses in this chapter, the value "1" on the dependent variable indicates that a person is food insecure. This first step in the transformation of the dependent variable makes it sensible to predict values between "0" and "1". Whereas a dichotomous variable has only two possible values ("0" or "1"), probabilities range from 0 to 1, with an infinite number of possible values between these two endpoints (0.2, 0.435, 0.58976, and so on). (See Chapter 5 for a review of probabilities.) So, instead of using the value on an independent variable to predict the value on a dependent variable, the value on an independent variable is used to predict the probability that the dependent variable has a "1" value. In mathematical notation, this is written as:

$$\Pr(y_i = 1 | x_i)$$

The “Pr” indicates that this is a probability. The “|” symbol means “given the condition” that is specified after the symbol. So, this equation is read as: the probability that the value on variable *y* (the dependent variable) is equal to “1”, given the value on variable *x* (the independent variable) for case *i*.

One way to proceed is to simply rework the linear regression prediction equation so that instead of predicting the values on the dependent variable ($\hat{y}$), it predicts the probabilities that the dependent variable has the value “1”. So, instead of using this prediction equation . . .

$$\hat{y} = a + bx$$

. . . researchers could use this prediction equation:

$$\Pr(\hat{y} = 1) = a + bx$$

This is a good first step, since probabilities have an infinite number of legitimate values between 0 and 1. But this approach still predicts probabilities below 0 or above 1, which are mathematically impossible.

Odds and Log Odds

odds Show the number of times that something occurs relative to the number of times that it does not occur.

The second step in transforming a dichotomous dependent variable so that it can be used in a regression is to move from probabilities to **odds**. Odds show the number of times something occurs relative to the number of times that it does not occur:

$$odds = \frac{number\ of\ times\ something\ occurs}{number\ of\ times\ something\ does\ not\ occur}$$

Transforming probabilities into odds is useful because odds range from 0 to positive infinity.

For example, if a course has 12 classes in a semester, and you attend 8 of them, that means that you do not attend 4 of them. Your odds of attending class are 8 divided by 4, or 2 (some people say “2 to 1” odds). A friend who is enrolled in the same course attends 6 classes and, therefore, does not attend 6 of them; that person’s odds of attending class are 6 divided by 6, or 1. (Some people say “one to one odds” or “even odds.”)

Odds can also be calculated for probabilities. For probabilities, the odds show the probability of something occurring relative to the probability of something not occurring:

$$odds = \frac{probability\ of\ something\ occurring}{probability\ of\ something\ not\ occurring}$$

For a dichotomous dependent variable, the probability of something not occurring (the denominator in the odds equation) is always equal to 1 minus the probability of something occuring. This is because there are only two possible values in a dichotomous variable, and thus the probability of having *either* value is equal to 1, or 100 per cent. For instance, if there's a 0.8 probability that a dichotomous variable has the value "1" for a specific case, then there's a 0.2 probability that it has the value "0" (1 – 0.8 = 0.2). Similarly, if there's a 0.3 probability that a dichotomous variable has the value "1" for a specific case, there's a 0.7 probability that it has the value "0" (1 – 0.3 = 0.7). So, for the probability that a dichotomous dependent variable has the value "1", the odds can be written as:

$$odds = \frac{\Pr\left(y_i = 1 | x_i\right)}{1 - \Pr\left(y_i = 1 | x_i\right)}$$

This equation is a bit unwieldy, so researchers often just use p_i to denote the probability that the value on y is equal to "1", given the value on x for that case. With this substitution, the odds are written as:

$$odds = \frac{p_i}{1 - p_i}$$

odds

Let's look at what happens when probabilities are transformed into odds. Remember that probabilities range from 0 to 1. The first column of Table 15.1 shows probabilities ranging from 0.01 to 0.99. The second and third columns of Table 15.1 show how each probability is transformed into odds. For a probability of 0, when something is guaranteed to not occur, the odds are also 0 (or 0 ÷ 100). For probabilities below 0.5, the odds are below 1 (i.e., something is unlikely to occur). A probability of 0.5 corresponds to an odds of 1, or even odds. For probabilities above 0.5, the odds are greater than 1 (i.e., something is likely to occur). And, in theory, a probability of 1, when something is guaranteed to occur, corresponds to an odds of positive infinity (or 100 ÷ 0). (Any number divided by 0 is equal to infinity.)

When the probability that a dichotomous dependent variable has the value "1" is transformed into odds, the values on the dependent variable can potentially range from "0" to positive infinity. But since regression also predicts values on the dependent variable that are below "0", one more step is needed in the transformation.

The third and final step in transforming a dichotomous dependent variable so that it can be used in a regression is a log transformation, which you learned about in Chapter 14. Recall that a log transformation represents a number (the odds) as the exponent of a common base number. When a variable is log-transformed, values less than "1" in the original variable become negative values in the transformed variable, the value "0" in the original variable becomes the value "1" in the transformed variable, and values greater than "1" in the original variable become positive values in the transformed variable. In Chapter 14, I described base 2 and base 10 log transformations.

In the natural sciences and mathematics, researchers often transform values using the **natural log** (or natural logarithm). A natural log transformation is just like a base 2 or a base 10 log transformation, only the number 2.71828 . . . is used as the common base number. The number 2.71828 . . . is called Euler's constant, after mathematician Leonhard Euler, and is denoted using the letter *e*. As with pi

natural log ($\log_e$) A logarithmic transformation using Euler's constant *e* (2.71828 . . .) as the common base number.

Table 15.1 Transforming Probabilities into Odds

Probability of Something Occurring	$\frac{p_i}{1 - p_i} =$	Odds of Something Occurring
0.01	$\frac{0.01}{0.99} =$	0.01
0.05	$\frac{0.05}{0.95} =$	0.05
0.1	$\frac{0.1}{0.9} =$	0.11
0.3	$\frac{0.3}{0.7} =$	0.43
0.5	$\frac{0.5}{0.5} =$	1.00
0.7	$\frac{0.7}{0.3} =$	2.33
0.9	$\frac{0.9}{0.1} =$	9.00
0.95	$\frac{0.95}{0.05} =$	19.00
0.99	$\frac{0.99}{0.01} =$	99.00

(π, 3.14159 . . .), the decimals of *e* continue on forever, and so 2.71828 . . . is only an approximate value. Natural log transformations have many mathematical properties that are particularly useful in statistical analysis, which is why they are sometimes used instead of log transformations with a more intuitive base number.

Table 15.2 illustrates what happens when odds are log-transformed using the common base *e*. When odds are transformed into the natural log ($\log_e$) of the odds, they are called **log odds**. After the natural log transformation, shown in the final column of Table 15.2, the values now include both positive and negative numbers. When there is less than a 0.5 probability of something occurring (i.e., it is unlikely to occur), the log odds are negative. When there is exactly a 0.5 probability of something occurring (i.e., even odds), the log odds are 0. And, when there is more than a 0.5 probability of something occurring (i.e., it is likely to occur), the log odds are positive. Theoretically, log odds range from negative infinity to positive infinity.

log odds The natural log of the odds of something occurring.

Table 15.2 Transforming Probabilities into Log Odds

Probability of Something Occurring	$\frac{p_i}{1 - p_i}$ = Odds	Odds = $e^?$	Natural Log ($\log_e$) of the Odds of Something Occurring
0.01	$\frac{0.01}{0.99} = 0.01$	$0.01 = e^{-4.60}$	−4.60
0.05	$\frac{0.05}{0.95} = 0.05$	$0.05 = e^{-2.94}$	−2.94
0.1	$\frac{0.1}{0.9} = 0.11$	$0.11 = e^{-2.20}$	−2.20
0.3	$\frac{0.3}{0.7} = 0.43$	$0.43 = e^{-0.85}$	−0.85
0.5	$\frac{0.5}{0.5} = 1$	$1 = e^{0}$	0.00
0.7	$\frac{0.7}{0.3} = 2.33$	$2.33 = e^{0.85}$	0.85
0.9	$\frac{0.9}{0.1} = 9$	$9 = e^{2.20}$	2.20
0.95	$\frac{0.95}{0.05} = 19$	$19 = e^{2.94}$	2.94
0.99	$\frac{0.99}{0.01} = 99$	$99 = e^{4.60}$	4.60

The notation that mathematicians use to indicate the process of finding the natural log of a number is *ln*. The *ln* function on a spreadsheet or calculator will show you that the natural log of 99 is 4.60 because 99 is equal to $e^{4.60}$. Similarly, the natural log of 19 is 2.94 because 19 is equal to $e^{2.94}$. So, the equation for the log odds is written as:

$$log\ odds = ln\left(\frac{p_i}{1 - p_i}\right)$$

log odds

Since the log odds range from negative infinity to positive infinity, the log odds of a dichotomous dependent variable can be substituted into a linear regression prediction equation without any problems. To review, the original dichotomous dependent variable is transformed so it can be used as the dependent variable in a regression in three steps:

1. First, the focus shifts to the probability that something will occur ($y = 1$).
2. Then, the probability is transformed into the odds that something will occur.
3. Finally, the odds are transformed into the log odds that something will occur.

The final, transformed variable is used as the dependent variable in a logistic regression prediction equation that is functionally equivalent to a linear regression prediction equation. This is the logistic regression prediction equation when only one independent variable is used:

logistic regression predictions (one independent variable)

$$ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = a + bx$$

Although logistic regression uses an approach that is similar to linear regression, there are several important differences between them:

- First, logistic regression is a type of non-linear regression. In Chapter 14, you learned about the difference between linear and non-linear transformations, and you learned that log transformations are non-linear. Because logistic regression incorporates a log transformation of the dependent variable, it does not predict a straight-line relationship between variables.
- Second, because of the transformation of the dependent variable, it's not possible to calculate the slope and constant coefficients of a logistic regression in the same way as for a linear regression. Instead, statistical software programs use a process called maximum likelihood estimation to calculate the constant and slope coefficients of a logistic regression.
- Finally, because of the transformation of the dependent variable, the interpretation of the regression coefficients changes. In the next section, I describe how to interpret the constant and slope coefficients produced by logistic regressions.

Interpreting Logistic Regression Coefficients

To illustrate how logistic regression coefficients are interpreted, let's return to the question of food security. In order to make the explanations in the remainder of this chapter easier to understand, I often refer to food insecurity as "going hungry." Overall, about 8 per cent of adults in Canada—or about 2.2 million people—are moderately or severely food insecure, that is, they go hungry. To start, let's use this percentage to calculate the odds of going hungry:

$$odds\ of\ going\ hungry = \frac{proportion\ of\ adults\ who\ go\ hungry}{proportion\ of\ adults\ who\ do\ not\ go\ hungry}$$

$$= \frac{8\%}{92\%} \text{ or } \frac{0.08}{0.92}$$

$$= 0.09$$

Since income is likely related to food security, let's also calculate the odds of going hungry for two separate groups: (1) people who have low income (less than $20,000 per year), and (2) people who have higher income ($20,000 or more per year). Among people who reported their income, about a third (30 per cent) have an annual income of less than $20,000. Among adults with annual incomes of less than $20,000, 15 per cent went hungry and 85 per cent did not go hungry. So, the odds of going hungry for people who have low income are calculated as:

$$\textit{odds of going hungry}_{\textit{low income}} = \frac{\begin{matrix}\textit{proportion of low income}\\ \textit{adults who go hungry}\end{matrix}}{\begin{matrix}\textit{proportion of low income}\\ \textit{adults who do not go hungry}\end{matrix}}$$

$$\textit{odds of going hungry}_{\textit{low income}} = \frac{15\%}{85\%} \text{ or } \frac{0.15}{0.85}$$

$$= 0.176$$

Among adults with an annual income of $20,000 or more, 5 per cent went hungry and 95 per cent did not go hungry. So, the odds of going hungry for people who have higher incomes are:

$$\textit{odds of going hungry}_{\textit{higher income}} = \frac{\begin{matrix}\textit{proportion of higher income}\\ \textit{adults who go hungry}\end{matrix}}{\begin{matrix}\textit{proportion of higher income}\\ \textit{adults who do not go hungry}\end{matrix}}$$

$$\textit{odds of going hungry}_{\textit{higher income}} = \frac{5\%}{95\%} \text{ or } \frac{0.05}{0.95}$$

$$= 0.053$$

In order to interpret logistic regression slope coefficients, it's crucial to understand the idea of an **odds ratio**. In Chapter 2 you learned about ratios, which show how the frequencies of two attributes compare directly to each other. Similarly, odds ratios show how the odds of something occurring in two different groups compare directly to each other. The odds of something occurring in the first group are expressed as a ratio of the odds of that same thing occurring in the second group. Just like odds, odds ratios range from 0 to infinity.

odds ratio Compares the odds of something occurring in two different groups; the odds of something occurring in the first group are expressed as a ratio of the odds of that thing occurring in the second group.

By using odds ratios, researchers can make fair comparisons between groups. For example, the odds of people with low income going hungry can be compared to the odds of people with higher income going hungry using the following ratio:

$$\textit{odds ratio} = \frac{\textit{odds of low income adults going hungry}}{\textit{odds of higher income adults going hungry}}$$

$$\begin{aligned} \textit{odds ratio} &= \frac{0.176}{0.053} \\ &= 3.32 \end{aligned}$$

An odds ratio of 1 indicates that the two groups have exactly the same odds of something occurring. An odds ratio greater than 1 indicates that the odds of something occurring for the group shown in the numerator of the ratio are higher than the odds of that thing occurring for the group shown in the denominator. An odds ratio less than 1 indicates that the odds of something occurring for the group shown in the numerator of the ratio are lower than the odds of that thing occurring for the group shown in the denominator. So, the odds ratio of 3.32 indicates that low income adults have higher odds of going hungry than higher income adults.

Researchers usually report *how much* higher (or lower) the odds of something occurring are for a specific group using percentages. Using percentages to report differences in odds makes sense because odds are non-linear. It's easiest to interpret odds ratios using a two-step process:

1. First, subtract 1 from the odds ratio. This accounts for the fact that an odds ratio of 1 indicates that both groups have the same odds of the thing occurring.
2. Second, multiply the remaining number by 100, in order to show the percentage difference in the odds.

Table 15.3 illustrates these two steps and how to interpret various odds ratios. The interpretation of odds ratios are typically framed as claims about how the odds of the group shown in the numerator of the ratio compare to the odds of the group shown in the denominator of the ratio. For instance, an odds ratio of 0.25 indicates that the group shown in the ratio's numerator has 75 per cent lower odds of something occurring than the group shown in the ratio's denominator. An odds ratio of 2.25 indicates that the group shown in the ratio's numerator has 125 per cent higher odds of something occurring than the group shown in the ratio's denominator.

Table 15.3 Interpreting Odds Ratios

Odds Ratio (OR)	Odds Ratio − 1 (OR − 1)	Percentage Difference in the Odds (OR − 1) (100)	Interpretation "The group in the numerator has ________ than the group in the denominator."
0.25	0.25 – 1 = −0.75	−75%	75 per cent lower odds of something occurring
0.50	0.50 – 1 = −0.50	−50%	50 per cent lower odds of something occurring
1.00	1.00 – 1 = 0	0%	Equal odds of something occurring
1.25	1.25 – 1 = 0.25	+25%	25 per cent higher odds of something occurring
1.50	1.50 – 1 = 0.50	+50%	50 per cent higher odds of something occurring
2.00	2.00 – 1 = 1.00	+100%	100 per cent higher odds of something occurring
2.25	2.25 – 1 = 1.25	+125%	125 per cent higher odds of something occurring

When the odds of adults with low income going hungry are compared to the odds of adults with higher income going hungry, the odds ratio is 3.32. As a result, we can assert that adults with low income have 232 per cent higher odds of going hungry than adults with higher income. I calculated this difference by subtracting 1 from 3.32 to get 2.32, and then multiplying by 100 to find the percentage difference (2.32 × 100 = 232 per cent difference).

Now that you understand the idea of an odds ratio, let's look at how they relate to logistic regression coefficients. Earlier in this chapter, you learned that in logistic regression the dichotomous dependent variable is transformed so that it captures the log odds of the probability that something will occur. Although this mathematical manipulation is useful for allowing researchers to use an approach that is similar to linear regression to make predictions about a dichotomous dependent variable, it makes the slope coefficients hard to interpret because they are expressed in log odds. For example, Table 15.4 shows the results of a logistic regression that uses low-income status, as a dummy variable, to predict whether or not people go hungry. As in a linear regression, the slope coefficient of the dummy variable shows how people who are in that group compare to people in the reference group. In this example, the slope coefficient of the "Has low income" dummy variable indicates that people with low income are predicted to have a log odds of going hungry that is 1.203 higher than people who do not have low income. The constant coefficient shows the prediction for people who have "0" values on all of the independent variables. In this example, the constant coefficient indicates that people who do not have low income (i.e., the "Has low income" dummy variable has a "0" value) are predicted to have a log odds of going hungry of –2.951. Although these results are technically correct, they aren't particularly meaningful.

To make logistic regression coefficients easier to interpret, researchers reverse the process of finding the natural log in order to turn the results back into odds. This is called finding the natural exponent of a number (*exp* or e^x). Finding the natural exponent of the log odds shown in the fourth column of Table 15.2 (using the *exp* or e^x function on a spreadsheet or calculator) transforms them back into the odds in the second column. So, the natural exponent of 4.60 is equal to 99 because $e^{4.60}$ is equal to 99. Similarly, the natural exponent of 2.94 is equal to 19 because $e^{2.94}$ is equal to 19.

In practice, the natural exponents of logistic regression slope coefficients are odds ratios. That is, they show how the odds of something occurring for the group captured in an independent dummy variable are predicted to compare to

Table 15.4 Results of a Logistic Regression with a Dummy Variable as an Independent Variable

Dependent variable: Is food insecure? (n = 48,787)

	Unstandardized Coefficient
Has low income (less than $20,000 a year)	1.203*
Constant	–2.951*

*Indicates that results are statistically significant at the p < 0.05 level.
Source: Author generated; Calculated using data from Statistics Canada, 2014.

Table 15.5 Results of a Logistic Regression with a Dummy Variable as an Independent Variable, with Odds Ratios

Dependent variable: Is food insecure? (n = 48,787)

	Unstandardized Coefficient	Odds Ratio
Has low income (less than $20,000 a year)	1.203*	3.33
Constant	−2.951*	0.05

*Indicates that results are statistically significant at the $p < 0.05$ level.
Source: Author generated; Calculated using data from Statistics Canada, 2014.

the odds of the same thing occurring for the reference group. Table 15.5 shows the same regression results as Table 15.4, but adds another column—labelled Odds Ratio—that shows the natural exponents of the unstandardized coefficients from Table 15.4. The odds ratio of the “Has low income” dummy variable is 3.33 because $e^{1.203}$ is equal to 3.33. So, these logistic regression results show that people with low income are predicted to have 233 per cent higher odds of going hungry than people with higher incomes (since 3.33 – 1 = 2.33, or 233 per cent). Notice that this matches the odds ratio calculated earlier in this chapter that compares the odds of adults with low income going hungry to the odds of adults with higher income going hungry. (That odds ratio was 3.32; the slight difference is due to rounding; the odds ratio calculated using percentages is equivalent to the odds ratio from this logistic regression because low-income status is the only independent variable used in the regression.)

If the odds ratio of a dummy variable is higher than 1, then the group captured in the dummy variable is predicted to be more likely than the reference group to have something occur. If the odds ratio of a dummy variable is exactly 1, then the group captured in the dummy variable is predicted to be just as likely as the reference group to have something occur. If the odds ratio of a dummy variable is lower than 1, then the group captured in the dummy variable is predicted to be less likely than the reference group to have something occur.

The interpretation of logistic regression constant coefficients is slightly different than the interpretation of logistic regression slope coefficients. Researchers still reverse the process of finding the natural log, and find the natural exponent of the unstandardized constant coefficient. But, for the constant coefficient, the result is just the odds—and not an odds ratio as it is for the slope coefficients. So, despite the column label, the value in the Odds Ratio column for the constant shows the odds of going hungry for people who have a “0” value on all of the independent variables. In Table 15.5, the constant coefficient of 0.05 shows that for higher income people (who have a “0” value on the “Has low income” dummy variable), the predicted odds of going hungry are 0.05. Notice that this also matches the odds of going hungry that we calculated for higher income adults (0.053) earlier in this chapter. (Again, the odds calculated using percentages are equivalent to the odds from this logistic regression because low-income status is the only independent variable used in the regression.)

Table 15.6 **Results of a Logistic Regression with a Ratio-Level Independent Variable, with Odds Ratios**

Dependent variable: Is food insecure? (n = 48,787)

	Unstandardized Coefficient	Odds Ratio
Age (in years)	−0.020*	0.98
Constant	−1.589*	0.20

*Indicates that results are statistically significant at the p < 0.05 level.
Source: Author generated; Calculated using data from Statistics Canada, 2014.

Up to this point, I have discussed odds ratios only for categorical independent variables because they are intuitively easier to understand. But ratio-level variables can also be used as independent variables in logistic regressions. Table 15.6 shows the results of a logistic regression that uses a ratio-level "Age" variable to predict whether people are likely to go hungry. Similar to linear regression, in a logistic regression the slope coefficient of a ratio-level independent variable shows how a one-unit increase in the independent variable is predicted to be associated with the dependent variable. In the logistic regression results shown in Table 15.6, the odds ratio of the "Age" variable is below 1, so as age increases, the odds of going hungry are predicted to be lower. Specifically, for each additional year older that people are, they are predicted to have 2 per cent lower odds of going hungry (since 0.98 – 1 = −0.02, or −2 per cent).

Because the "Age" variable used in this logistic regression is not centred, the constant shows the odds of going hungry for people who are 0 years old; thus, it makes little sense to discuss this result. Most of the time, researchers do not discuss the odds of the constant coefficient when they report the results of a logistic regression. This is also why many statistical software programs print the constant at the bottom of logistic regression results.

The two logistic regressions I have shown so far have only used a single independent variable. But the main advantage of using regression—including logistic regression—is the ability to predict the unique relationship between an independent variable and a dependent variable while controlling for a series of other variables. For instance, the odds of going hungry are likely related to other things besides income, and a more complex logistic regression model can help to identify which characteristics are the strongest predictors of food insecurity.

Table 15.7 shows the results of a logistic regression that uses age, sex/gender, annual personal income, and region of residence to predict the relative odds of going hungry. The annual personal income dummy variables now capture four levels of income, instead of just distinguishing between people who have low income and those who do not. The odds ratios show how people's income is related to their odds of going hungry. Compared to people with annual incomes of $80,000 or more, people with annual incomes of less than $20,000 are predicted to have 2,229 per cent higher odds of going hungry, people with annual incomes of $20,000 to $39,999 are predicted to have 1,252 per cent higher odds of going hungry, people

Table 15.7 **Results of a Logistic Regression, with Odds Ratios**

Dependent variable: Is food insecure? (n = 48,787)

	Unstandardized Coefficient	Odds Ratio
Annual personal income (ref: $80,000 or more)		
Less than $20,000	3.148*	23.29
$20,000 to $39,999	2.604*	13.52
$40,000 to $59,999	1.909*	6.75
$60,000 to $79,999	1.042*	2.84
Region of residence (ref: Ontario)		
Atlantic Canada	0.213*	1.24
Quebec	0.059	1.06
Prairies	0.106*	1.11
British Columbia & the Territories	0.136*	1.15
Age (in years)	–0.014*	0.99
Women	–0.014	0.99
Constant	–4.386*	0.01

*Indicates that results are statistically significant at the p < 0.05 level.
Source: Author generated; Calculated using data from Statistics Canada, 2014.

with annual incomes of $40,000 to $59,999 are predicted to have 575 per cent higher odds of going hungry, and people with annual incomes of $60,000 to $79,999 are predicted to have 184 per cent higher odds of going hungry, after controlling for region of residence, age, and sex/gender.

The estimated odds of going hungry also vary depending on where people live in Canada. Compared to people who live in Ontario, people living in Atlantic Canada are predicted to have 24 per cent higher odds of going hungry, people living in the Prairies are predicted to have 11 per cent higher odds of going hungry, and people living in British Columbia and the three territories are predicted to have 15 per cent higher odds of going hungry, after controlling for annual personal income, age, and sex/gender. (The three territories are grouped with British Columbia because there are few cases from the territories.) Age appears to have a weaker relationship with the odds of going hungry. After controlling for annual personal income, region of residence, and sex/gender, each one-year increase in age is associated with having only 1 per cent lower odds of going hungry.

The logistic regression shown in Table 15.7 begins to introduce more variables into the examination of adults' food insecurity. As with linear regression, researchers strive to build logistic regression models that capture the wide range of characteristics and social processes that might be related to a dependent variable. The techniques for manipulating independent variables that you learned about in Chapter 14—interactions, quadratics, and transformations—can also be used in logistic regressions, although interpreting the coefficients becomes increasingly complex.

Standardized Slope Coefficients

So far, none of the logistic regression results have included standardized slope coefficients. Although some statistical software programs produce standardized slope coefficients for logistic regressions, others (such as SPSS) do not. As you learned in Chapter 12, though, a standardized slope coefficient can be calculated using the unstandardized slope coefficient and the standard deviations of both the independent and dependent variables. Recall that the formula for calculating a standardized slope coefficient is:

$$\beta_x = b_x\left(\frac{s_x}{s_y}\right)$$

This same formula can be used to calculate standardized slope coefficients for logistic regressions, using the predicted difference in the log odds (shown in the "Unstandardized Coefficient" [b] column) and the standard deviations of the independent and dependent variables. Table 15.8 shows the standardized slope coefficients of the independent variables used in the logistic regression in Table 15.7. The standard deviation of the dependent variable and each of the independent variables was obtained using statistical software, and then used to calculate each of the standardized slope coefficients.

Table 15.8 Standardized Slope Coefficients for the Logistic Regression Shown in Table 15.7

Dependent variable: Is food insecure? (n = 48,787)

	Unstandardized Coefficient (b_x)	Standard Deviation of the Independent Variable (s_x)	Standard Deviation of the Dependent Variable (s_y)	Standardized Coefficient (β_x)
Annual personal income (ref: $80,000 or more)				
Less than $20,000	3.148*	0.458	0.270	5.35
$20,000 to $39,999	2.604*	0.445	0.270	4.30
$40,000 to $59,999	1.909*	0.390	0.270	2.76
$60,000 to $79,999	1.042*	0.313	0.270	1.21
Region of residence (ref: Ontario)				
Atlantic Canada	0.213*	0.258	0.270	0.20
Quebec	0.059	0.428	0.270	0.09
Prairies	0.106*	0.377	0.270	0.15
British Columbia & the Territories	0.136*	0.347	0.270	0.18
Age (in years)	−0.014*	17.578	0.270	−0.91
Women	−0.014	0.500	0.270	−0.03

*Indicates that results are statistically significant at the $p < 0.05$ level.

Source: Author generated; Calculated using data from Statistics Canada, 2014.

The standardized slope coefficients of logistic regressions are interpreted in the same way as those of linear regressions. Because the units of measurement aren't easily interpretable, researchers typically just identify the independent variable with the highest absolute standardized slope coefficient (regardless of whether it is positive or negative) as that which has the strongest relationship with the dependent variable. Similarly, researchers can identify the independent variable with the lowest absolute standardized slope coefficient as that which has the weakest relationship with the dependent variable. In this example, the annual personal income dummy variables (collectively) have the highest standardized slope coefficients. Thus, among the independent variables in this logistic regression, annual personal income has the strongest relationship with food insecurity. The "Women" dummy variable has the smallest standardized slope coefficient; thus, among the independent variables in this logistic regression, sex/gender has the weakest relationship with food insecurity.

Statistical Significance Tests and Confidence Intervals

As you might expect, researchers are also interested in assessing the reliability, or statistical significance, of logistic regression results. Logistic regression relies on a version of a chi-square test of statistical significance, called a Wald chi-square, to assess the likelihood of randomly selecting a sample with the observed relationship, or one of greater magnitude, if no relationship exists between an independent variable and the dependent variable in the larger population. In the context of logistic regression, significance tests show the likelihood of randomly selecting a sample with the observed log odds (or larger log odds), if the log odds of the relationship between an independent variable and the dependent variable are actually 0 (corresponding to a probability of 0.5 or odds of 1) in the population. The p-values produced by the Wald chi-square test are interpreted in the same way as all other p-values.

Notice that I did not discuss the odds ratios of the "Quebec" dummy variable in the description of the logistic regression results in Table 15.7. Since the unstandardized slope coefficient of the "Quebec" dummy variable is not statistically significant (it has a p-value greater than 0.05), we are not confident that adults in the Quebec population have different odds of going hungry than adults in the Ontario population (the reference group), after controlling for the other variables in the regression. Similarly, since the unstandardized slope coefficient of the "Women" dummy variable is not statistically significant, we are not confident that—in the Canadian population—the odds of women of going hungry are any different than the odds of men of going hungry, after controlling for annual personal income, region of residence, and age.

Similar to linear regression, researchers sometimes present the 95 per cent confidence intervals for logistic regression coefficients. Typically, researchers present the confidence intervals for the odds ratio of each independent variable, as opposed to the confidence intervals for the log odds. The 95 per cent confidence interval for an odds ratio shows the range that the odds ratio capturing each relationship is likely to be within in the population that the sample was selected from, with 95 per cent confidence. If the 95 per cent confidence interval for the odds ratio of a categorical independent variable (a dummy variable) overlaps with 1, researchers cannot be confident that, in the population, the group captured in the dummy variable has different odds of the outcome captured in the dependent variable than the reference group. Similarly, if the 95 per cent confidence interval for the odds ratio of a ratio-level independent variable overlaps with 1, researchers cannot be confident that, in the population, a one-unit increase in the independent variable is associated with any change in the odds of the outcome captured in the dependent variable.

Table 15.9 shows 95 per cent confidence intervals for the odds ratios of the independent variables used in the logistic regression shown in Table 15.7. The 95 per cent confidence intervals for the odds ratios are used to illustrate the amount of uncertainty in researchers' estimates; this uncertainty is the result of using information from a random sample of cases to make predictions, instead of using information from the entire population. For instance, researchers can assert that they are 95 per cent confident that the odds of going hungry in the adult population of Atlantic Canada are somewhere between 9 to 41 per cent higher than the odds of going hungry in the Ontario adult population. Alternatively, a researcher might report that the odds of going hungry in the adult population of British Columbia and the Territories are 15 per cent higher (95% CI: 3%–27%) than the odds of going hungry in the Ontario adult population.

Table 15.9 Results of a Logistic Regression, with Odds Ratios and 95 Per Cent Confidence Intervals

Dependent variable: Is food insecure? (n = 48,787)

	Unstandardized Coefficient	Odds Ratio	95% Confidence Interval for the Odds Ratio	
			Lower Bound	Upper Bound
Annual personal income (ref: $80,000 or more)				
Less than $20,000	3.148*	23.29	17.24	31.47
$20,000 to $39,999	2.604*	13.52	9.99	18.30
$40,000 to $59,999	1.909*	6.75	4.94	9.22
$60,000 to $79,999	1.042*	2.84	1.99	4.03
Region of residence (ref: Ontario)				
Atlantic Canada	0.213*	1.24	1.09	1.41
Quebec	0.059	1.06	0.97	1.16
Prairies	0.106*	1.11	1.00	1.23
British Columbia & the Territories	0.136*	1.15	1.03	1.27
Age (in years)	–0.014*	0.99	0.98	0.99
Women	–0.014	0.99	0.92	1.06
Constant	–4.386*	0.01		

*Indicates that results are statistically significant at the p < 0.05 level.

Source: Author generated; Calculated using data from Statistics Canada, 2014.

Notice that the 95 per cent confidence interval for the "Quebec" dummy variable shows that—in the population of adults—the odds ratio is likely to be between 0.97 and 1.16. In other words, compared to the Ontario population, the odds of going hungry could be 3 per cent lower in the Quebec population *or* 16 per cent higher in the Quebec population (or anywhere in between). Thus, the odds of going hungry could plausibly be exactly the same for adults in the Quebec population and in the Ontario population. In this example, the 95 per cent confidence interval mirrors the results of the statistical significance test, which shows that there is not a statistically significant difference between Quebec adults' odds of going hungry and Ontario adults' odds of going hungry (Ontario is the reference group). The upper and lower bounds of the 95 per cent confidence interval for the odds ratio of the "Women" dummy variable also overlap with 1, mirroring the results of the statistical significance test. Often, when the upper and lower bounds of the 95 per cent confidence interval overlap with 1, the result is not statistically significant (but not always). In general, the width of the 95 per cent confidence intervals for odds ratios provide information about how precise logistic regression predictions are.

Statistics in Use

Which Low-Income Families Use Food Banks?

Original research: Loopstra, Rachel, and Valerie Tarasuk. 2012. "The Relationship between Food Banks and Household Food Insecurity among Low-Income Toronto Families." Canadian Public Policy *38 (4): 497–514.*

Researchers know that many people who are food insecure do not use food banks. A team of researchers wanted to find out which low-income families were more likely to use food banks and which were less likely to do so (Loopstra and Tarasuk 2012). The researchers sampled low-income families from 12 randomly selected high-poverty areas in Toronto; in order to be included in the sample, families needed to rent their home, have at least one child aged 18 or younger, and have enough English-language skills to complete an interview. Information was collected from the household member who was primarily responsible for shopping and food management.

The researchers asked respondents whether or not their family had used a food bank during the past 12 months. A logistic regression was used to determine how the odds of food bank use are associated with various household characteristics. The number and percentage of households with each characteristic, grouped by whether or not they used a food bank in the past 12 months, are shown in the first two columns of Table 15.10. The odds ratios (OR) and 95 per cent confidence intervals produced by a logistic regression are shown in the final column of Table 15.10. For each household characteristic, the row with an odds ratio of 1 (in the final column) indicates the reference group. In addition to the

Table 15.10 Household Characteristics in Relation to Food Bank Use (N = 371)

Household Characteristics	Did Not Use Food Bank (n = 287)	Used Food Bank (n = 84)	Adjusted OR[a] (95% CI)
Food security status, *n* (%)			
Food secure	88 (94)	6 (6)	1.00
Marginally food insecure	42 (89)	5 (11)	1.48 (0.30–7.22)
Moderately food insecure	89 (75)	29 (25)	3.21 (1.26–8.18)
Severely food insecure	68 (61)	44 (39)	3.75 (1.18–11.90)
12-month income[b] (mean ± SE)	28,339 ± 631.90	20,843 ± 1,180.78	1.19 (1.11–1.26)[c]
Received welfare, *n* (%)			
No	209 (89)	25 (11)	1.0
Yes	78 (57)	59 (43)	3.19 (1.52–6.70)
Immigrated ≤ 5 years ago, *n* (%)			
No	232 (75)	79 (25)	1.0
Yes	55 (92)	5 (8)	0.37 (0.16–0.85)
Household type, *n* (%)			
Two-parent or lone father	142 (87)	21 (13)	1.0
Lone mother	145 (70)	63 (30)	0.59 (0.25–1.39)
Baseline education of respondent, *n* (%)			
Some or completed post-secondary	142 (86)	24 (14)	1.0
High school	97 (78)	28 (22)	0.86 (0.29–2.57)
Less than high school	48 (60)	32 (40)	1.33 (0.59–3.01)
Have children ≤ 3 years old, *n* (%)			
No	206 (77)	62 (23)	1.0
Yes	81 (79)	22 (21)	0.90 (0.48–1.66)

Notes:

[a] Logistic regression model adjusted for variables in table, number of adults and children in household, receipt of housing subsidy, and clustering effect of neighbourhood.

[b] Income means adjusted for number of adults and children in household.

[c] Income OR refers to a $2,000 decrease in income.

Source: Calculated by authors.

Source: Loopstra and Tarasuk 2012, 503. Reprinted with permission from University of Toronto Press (www.utpjournals.com).

Continued

characteristics listed, the regression also controls for the number of adults and children living in the household, whether or not the household received a housing subsidy, and neighbourhood effects. (See Table 15.10, note [a].)

The researchers assert that "the odds of using a food bank in the past 12 months . . . increased with the severity of food insecurity" (Loopstra and Tarasuk 2012, 501). As households have progressively higher levels of food insecurity, the predicted odds ratios increase: from 1.48 for marginally food-insecure households, to 3.21 for moderately food-insecure households, to 3.75 for severely food-insecure households (all in comparison to food-secure households). The logistic regression results also show that the odds of food bank use are 16 per cent higher for every $2,000 decrease in household income, adjusted for household size. (Note [c] on the table indicates that the odds ratio refers to a $2,000 decrease.) Households who receive social assistance (welfare) have 219 per cent higher odds of using food banks than those who do not. Notably, recent immigrants are less likely to use food banks than more settled immigrants or the Canadian-born, since the odds ratio of the recent immigrant ("Yes") dummy variable is less than 1.

The 95 per cent confidence interval for each odds ratio is shown in parentheses following the odds ratio. The wide 95 per cent confidence intervals for the odds ratios of the household type and education variables make it difficult to draw any conclusions about how these characteristics are related to food bank use. Similarly, no conclusions can be made about how having a child aged three or under in the household is related to food bank use.

Some study participants indicated that they do not use food banks because they don't provide suitable food, they are able to manage without using them, they associate food bank use with degrading feelings, or they don't feel like food banks are meant for them. Other participants cited barriers to food bank use related to access (such as not being able to go during open hours or not being able to document their financial need) and not having enough information about them.

Based on this research, Loopstra and Tarasuk conclude that most food-insecure families are not using food banks and that, among those families who are using food banks, they do not do much to alleviate food insecurity. In response, they make a series of policy recommendations for improving both access to food banks and the resources that they are able to provide.

Calculating Predicted Probabilities

Researchers often use logistic regression results to make claims about how the probability that something will occur is different for people with different characteristics. Recall that the *probability* that something will occur is not the same as the *odds* that something will occur. The first and last columns of Table 15.1 illustrate the difference between probabilities and odds. Fortunately, the predicted probability that the outcome captured in the dependent variable will occur for a specific type of person (or case) can be calculated using unstandardized logistic regression coefficients. This process is the equivalent of using the linear regression prediction equation to predict the value on the dependent variable for a specific type of person (or case).

Calculating the predicted probability that the outcome captured in the dependent variable will occur for a specific case (using logistic regression coefficients) is a two-step process. The first step is to calculate a z-value (which is different than a z-score). This step is very similar to calculating a linear regression prediction. The type of person (or case) that the prediction is being made for is specified by modifying the value on each independent variable in the equation. The formula for calculating a z-value should look familiar:

$$z = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + \cdots$$

z-value

Notice that the log odds (shown in the "Unstandardized Coefficient" column of logistic regression results) are used to calculate the z-value, and not the odds ratios.

The second step is to use the z-value to calculate the predicted probability of the outcome captured in the dependent variable occuring. This additional step accounts for the transformations that are made to the dependent variable in a logistic regression. The predicted probability ($\hat{p}$) is calculated as:

$$\hat{p} = \frac{e^z}{1 + e^z}$$

predicted probability

In this formula, *e* represents Euler's constant (2.71828 . . .), which was described earlier in this chapter.

Let's use the logistic regression results in Table 15.5 to calculate the predicted probability of going hungry. The regression uses a single categorical independent variable: low-income status. The unstandardized constant coefficient (*a*) is –2.951, and the unstandardized slope coefficient (*b*) is 1.203. These coefficients are substituted into the z-value formula:

$$z = a + b_1 x_1$$
$$z = -2.951 + 1.203(\mathit{low\ income})$$

For people who do not have low income, the z-value is:

$$\begin{aligned} z &= -2.951 + 1.203(\mathit{low\ income}) \\ &= -2.951 + 1.203(0) \\ &= -2.951 \end{aligned}$$

For people who do have low income, the z-value is:

$$\begin{aligned} z &= -2.951 + 1.203(\mathit{low\ income}) \\ &= -2.951 + 1.203(1) \\ &= -1.748 \end{aligned}$$

Let's proceed to the second step, and substitute each of these z-values into the equation to find the predicted probability of going hungry. For people who do not have low income, the predicted probability of going hungry is:

$$\begin{aligned} \hat{p} &= \frac{e^z}{1+e^z} \\ \hat{p} &= \frac{e^{-2.951}}{1+e^{-2.951}} \\ &= \frac{0.052}{1+0.052} \\ &= \frac{0.052}{1.052} \\ &= 0.05 \end{aligned}$$

For people who have low income, the predicted probability of going hungry is:

$$\begin{aligned} \hat{p} &= \frac{e^z}{1+e^z} \\ \hat{p} &= \frac{e^{-1.748}}{1+e^{-1.748}} \\ &= \frac{0.174}{1+0.174} \\ &= \frac{0.174}{1.174} \\ &= 0.15 \end{aligned}$$

So, people who do not have low income have a 0.05 probability, or a 5 per cent chance, of going hungry; and people who have low income have a 0.15 probability, or a 15 per cent chance, of going hungry. Notice that these probabilities correspond exactly to the percentage of adults in each group who are food insecure, reported earlier in this chapter. This result only occurs when a logistic regression uses a single, categorical independent variable. Once logistic regression models become more complex, the predicted probabilities no longer correspond to the group percentages because the predictions control for the other independent variables in the regression.

When a logistic regression uses a ratio-level independent variable, researchers usually calculate the predicted probabilities for a series of typical values and then graph them in order to show the predicted relationship. The logistic regression results shown in Table 15.6 predict the odds of going hungry, using a ratio-level "Age" variable as the independent variable. The regression coefficients in Table 15.6 are substituted into the z-value equation:

$$z = a + b_1 x_1$$
$$z = -1.589 + (-0.020)(age)$$

To better understand the relationship between age and the probability of going hungry, a z-value is calculated for people who are aged 15, aged 20, aged 25, and so on all the way up to age 80. (The oldest people in the sample are 80.) Then, predicted probabilities are calculated for people at each age, and the results are graphed to show the general pattern. Figure 15.2 shows that—in general—as age increases, the predicted probability of going hungry decreases. Notice that the predicted relationship between age and the probability of going hungry is curvilinear. Earlier in this chapter, you learned that logistic regression is a type of non-linear regression, and, thus, it does not predict a straight-line relationship.

This same two-step process of calculating z-values and predicted probabilities can also be used when logistic regressions use more than one independent variable. In these situations, however, researchers usually choose only one or two variables to highlight, varying the values of those variables while holding the other independent variables constant. For instance, the logistic regression shown in Table 15.7 uses four independent variables (annual personal income, region of residence, age, and sex/gender) and has 10 slope coefficients and a constant coefficient. It's simply unwieldy to display the predicted probabilities for every possible combination of characteristics. Since I am primarily interested in the relationship between income and the probability of going hungry, I decided to calculate and plot the predicted probabilities of going hungry for people in three of the five personal income groups: the lowest income group (who have annual incomes of less than $20,000),

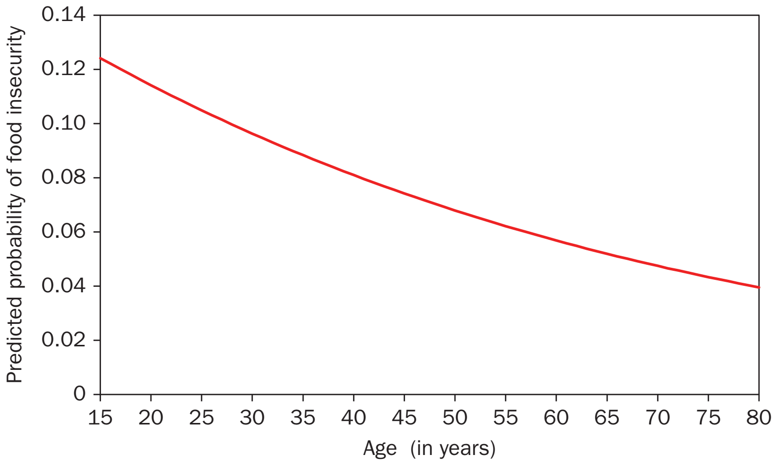


Figure 15.2 **Predicted Probability of Food Insecurity, by Age, Calculated from the Logistic Regression Results in Table 15.6**

Source: Author generated; Calculated using data from Statistics Canada, 2014.

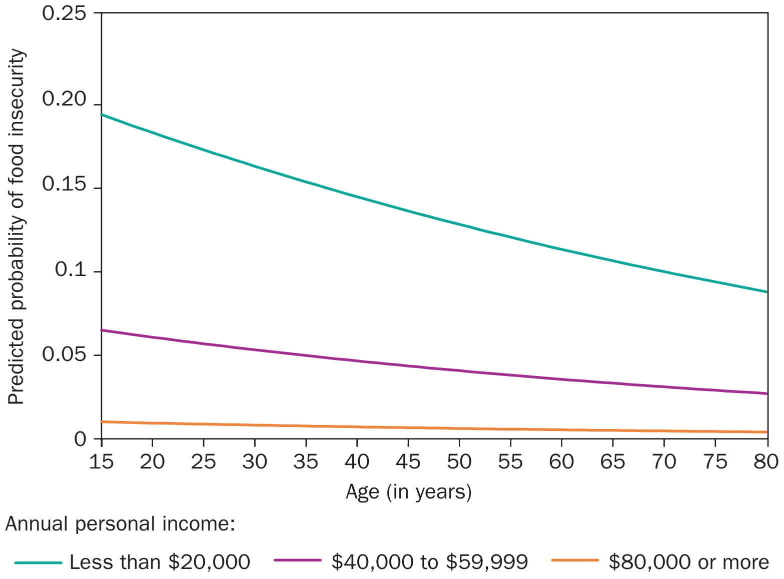


Figure 15.3 **Predicted Probability of Food Insecurity for Men Living in Ontario, by Age, for Three Annual Personal Income Groups, Calculated from the Logistic Regression Results in Table 15.7**

Source: Author generated; Calculated using data from Statistics Canada, 2014.

the middle income group (who have annual incomes of $40,000 to $59,999), and the highest income group (who have annual incomes of $80,000 or more). Since age also has a statistically significant slope coefficient, I include predictions for people at various ages in each of the three income groups. But to avoid further complexity, I only show predictions for people who live in Ontario (since the largest proportion of the Canadian population lives in Ontario). Similarly, I only show predictions for men because we are not confident that the odds of going hungry are different for women than for men.

Figure 15.3 shows the predicted probability of going hungry for men living in Ontario, with three different levels of annual personal income and across a range of ages. The graph shows that among men with the highest personal incomes, there is a very weak relationship between age and the probability of going hungry. Indeed, the probability of going hungry is very low for men of all ages who have an annual personal income of $80,000 or more. Men with the lowest personal incomes, who have an annual income of less than $20,000, have a much higher probability of going hungry overall. In addition, the relationship between age and the probability of going hungry is much stronger among men with the lowest personal incomes.

When researchers report the results of logistic regressions, they sometimes present the estimated odds ratios; sometimes, predicted probabilities; and other times, both. In order to ensure that you understand what logistic regression results are showing, be sure to note whether odds ratios or probabilities are being reported.

Step-by-Step: Predicted Probabilities (Logistic Regression)

Z-values

$$z = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + \cdots$$

Step 1 Identify the log odds of each independent variable (b_1, b_2, b_3, . . .) in a logistic regression. If only odds ratios are available, calculate the natural log (*ln*) of the odds ratio to obtain the log odds.

Step 2 Identify the log odds of the constant (a) in a logistic regression. If only the odds are available, calculate the natural log (*ln*) of the odds to obtain the log odds.

Step 3 Determine the values of the independent variable (x_1, x_2, x_3, . . .) for the specific case for which you want to predict the probability that the outcome captured in the dependent variable will occur.

Step 4 For *each independent variable* in the logistic regression, multiply the log odds associated with the variable (from Step 1) by the value on the variable for the specific case you want to make a prediction for (from Step 3).

Step 5 Add together the results of Step 4 for each independent variable.

Step 6 Add the log odds of the constant (from Step 2) to the result of Step 5 to find the z-value for the specific case.

Predicted Probability

$$\hat{p} = \frac{e^z}{1 + e^z}$$

Step 7 Find the natural exponent of the z-value for the specific case (from Step 6). Do this using the e^x function on a calculator, or the EXP() function in a spreadsheet.

Step 8 Add 1 to the result of Step 7 to find the denominator of the predicted probability equation.

Step 9 Divide the result of Step 7 by the result of Step 8 to find the predicted probability that the outcome captured in the dependent variable will occur for the specific case.

Assessing Model Fit for Logistic Regressions

For linear regressions, researchers use the R^2 or the adjusted R^2 to determine how much of the variation in the dependent variable can be explained by the independent variables. The R^2 and the adjusted R^2 provide an overall assessment of how well a linear regression model fits the observed data. Unfortunately, there is no exact equivalent to R^2 for logistic regressions. But many statisticians have tried to develop a way to assess the overall fit of a logistic regression that is similar to the R^2 of a linear regression. These are called pseudo-R^2s because, although they are conceptually similar to R^2, they are not exactly the same. A pseudo-R^2 typically reports how much the predictions made by a logistic regression, with one or more independent variables, are an improvement over a null model. You might recall from Chapter 11 that a null model is one that does not use any independent variables.

Nagelkerke's R^2 A pseudo-R^2 that is commonly used to assess the overall fit of a logistic regression.

The most common pseudo-R^2 that researchers report for logistic regressions is called **Nagelkerke's R^2**, which is popular because it ranges from 0 to 1, exactly like the R^2 of a linear regression. Also like R^2, the larger Nagelkerke's R^2 is, the better a logistic regression model fits the observed data. However, because of how Nagelkerke's R^2 is calculated, it can't be used to compare logistic regressions that use different cases or different dependent variables. Instead, Nagelkerke's R^2 is most useful for comparing nested logistic regressions that use the same cases to predict the same dependent variable. You learned about nested regressions and using blocks to group independent variables in Chapter 13; the same approach can be used in logistic regressions. If Nagelkerke's R^2 increases substantially when independent variables are added to a nested logistic regression, this indicates that the variables that were added notably improve the fit of the model. If Nagelkerke's R^2 does not increase much when independent variables are added to a nested logistic regression, this indicates that the variables that were added do not particularly improve the fit of the model.

To illustrate, I divided the logistic regression shown in Table 15.7 into a nested regression with three blocks: the first block uses the dummy variables that capture annual personal income, the second block adds the dummy variables that capture region of residence, and the third block adds the "Age" variable and the "Women" dummy variable. The Nagelkerke's R^2 of the first model, which predicts the odds of going hungry using annual personal income alone, is 0.098. In the second model, which incorporates region of residence into the prediction, Nagelkerke's R^2 is 0.099; there is only a very small increase. In the third model, which adds age and sex/gender as predictors (and corresponds to the logistic regression in Table 15.7), Nagelkerke's R^2 increases to 0.109. These results indicate that accounting for personal income substantially improves the logistic regression predictions, but accounting for region of residence does not improve the model much more. Taking age and sex/gender into account is associated with a slight increase in Nagelkerke's R^2, but there is room for this logistic regression model to be improved much further.

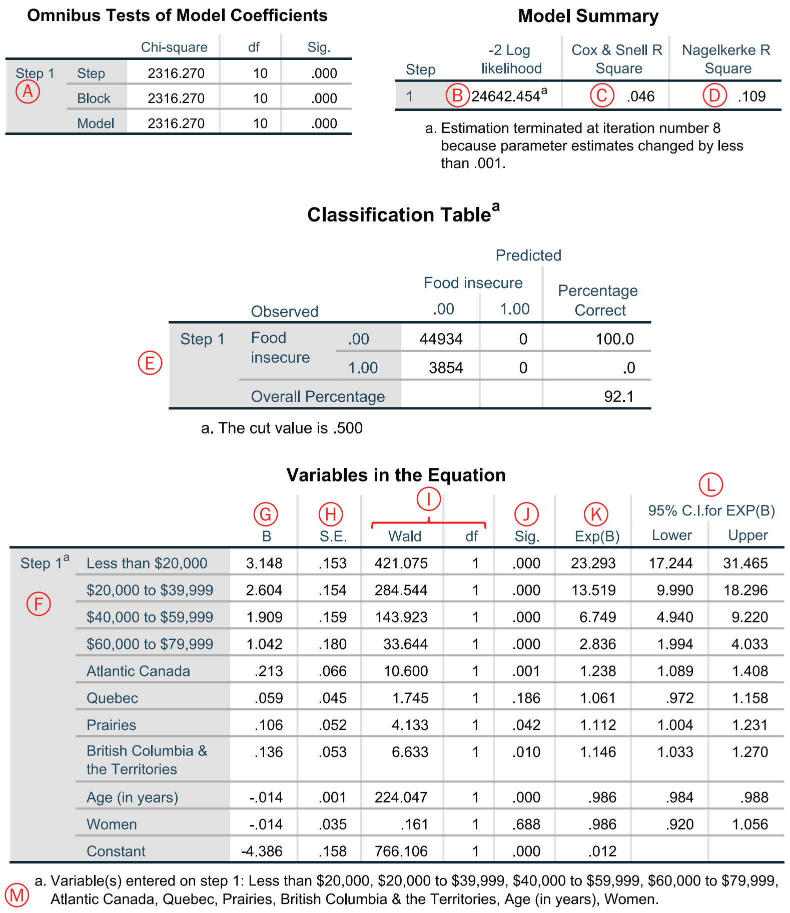
Many of the other strategies that researchers use to assess how well a linear regression model fits the data do not transfer easily to logistic regression models. For instance, there is no simple equivalent to analyzing regression residuals in the context of logistic regression. Similarly, the logistic regression procedures in some statistical software do not produce collinearity statistics. This does not mean that collinearity is not a concern in logistic regression, however, and researchers should still be alert to any strong correlations between independent variables. An investigation of the bivariate relationships between the independent variables, and between each independent variable and the dependent variable, should be the starting point of any regression analyses—regardless of whether researchers are using linear regression or logistic regression techniques.

How Does It Look in SPSS?

Logistic Regression

When the independent variables are entered in a single block and the 95 per cent “CI for exp(b)” option is selected, the Binary Logistic Regression procedure, produces results that look like those in Image 15.1. A similar sequence of results is printed for the null model—a model without any predictors—with the label “Block 0: Beginning Block” (not shown). The results shown here are labelled “Block 1.”

A. Independent variables are entered into a logistic regression in blocks. This regression only has a single block/step, labelled “Step 1,” and thus the results of the step, the block, and the model in the “Omnibus Tests of Model Coefficients” table are identical. A chi-square test is used to assess whether or not, as a group, the independent variables are likely to be related to the dependent variable in the population. The degrees of freedom


A
B
C
D
E
F
G
H
I
J
K
L
M

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	2316.270	10	.000
	Block	2316.270	10	.000
	Model	2316.270	10	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	24642.454[a]	.046	.109

a. Estimation terminated at iteration number 8 because parameter estimates changed by less than .001.

Classification Table[a]

	Observed		Predicted		
			Food insecure		Percentage Correct
			.00	1.00	
Step 1	Food insecure	.00	44934	0	100.0
		1.00	3854	0	.0
	Overall Percentage				92.1

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Less than $20,000	3.148	.153	421.075	1	.000	23.293	17.244	31.465
	$20,000 to $39,999	2.604	.154	284.544	1	.000	13.519	9.990	18.296
	$40,000 to $59,999	1.909	.159	143.923	1	.000	6.749	4.940	9.220
	$60,000 to $79,999	1.042	.180	33.644	1	.000	2.836	1.994	4.033
	Atlantic Canada	.213	.066	10.600	1	.001	1.238	1.089	1.408
	Quebec	.059	.045	1.745	1	.186	1.061	.972	1.158
	Prairies	.106	.052	4.133	1	.042	1.112	1.004	1.231
	British Columbia & the Territories	.136	.053	6.633	1	.010	1.146	1.033	1.270
	Age (in years)	-.014	.001	224.047	1	.000	.986	.984	.988
	Women	-.014	.035	.161	1	.688	.986	.920	1.056
	Constant	-4.386	.158	766.106	1	.000	.012		

a. Variable(s) entered on step 1: Less than $20,000, $20,000 to $39,999, $40,000 to $59,999, $60,000 to $79,999, Atlantic Canada, Quebec, Prairies, British Columbia & the Territories, Age (in years), Women.

Image 15.1 An SPSS Logistic Regression

Continued

are equal to the number of independent variables in the regression. These results are not usually reported.

B. The "–2 Log Likelihood" statistic is used to compare nested logistic regression models. Lower values indicate a better-fitting model.

C. The "Cox & Snell R Square" is a pseudo-R^2. Higher values indicate a better-fitting model.

D. The "Nagelkerke R Square" is a pseudo-R^2. Higher values indicate a better-fitting model.

E. The "Classification Table" shows how the observed values on the dependent variable compare to the values predicted by the logistic regression. This regression predicts that every case is in the "No" group (not food insecure). The "Percentage Correct" column shows that this regression makes the correct prediction for 92 per cent of cases.

F. The independent variables are listed in rows. The row for the constant is always at the bottom and is typically not discussed.

G. The "B" column shows the log odds of each independent variable and the constant (the unstandardized coefficients). The log odds are used to calculate the odds ratios and the predicted probabilities.

H. This column shows the standard error of the log odds of each independent variable and the constant. The standard error is used to calculate the confidence intervals.

I. These columns show the Wald chi-square statistic (which is slightly different than the Pearson chi-square statistic you learned about in Chapter 9) and the degrees of freedom of the distribution it is evaluated against.

J. The "Sig." column shows the p-value associated with each Wald chi-square statistic. The results are interpreted in the same way as all other p-values: they show the likelihood of randomly selecting a sample with the observed relationship (or one of a greater magnitude), if no relationship exists between the independent variable and the dependent variable in the population. For example, the p-values of the four annual personal income dummy variables indicate that there is a less than 0.1 per cent chance of selecting this sample from a population in which there is no relationship between annual personal income and food security. (All are $p < 0.001$.)

K. The "Exp(B)" column shows the odds ratio associated with each independent variable and the odds associated with the constant. The numbers in this column are the natural exponents of the log odds in the "B" column. For example, these results show that people living in Atlantic Canada have 24 per cent higher odds of being food insecure than people living in Ontario (the reference group); people living in the Prairies have 11 per cent higher odds of being food insecure than people living in Ontario.

L. These columns show the lower and upper bounds of the 95 per cent confidence interval for each odds ratio. The width of the confidence interval shows how precise the logistic regression estimates are. In the population, people living in Atlantic Canada are estimated to have between 9 and 41 per cent higher odds of being food insecure than people living in Ontario.

M. The footnote below this table lists the independent variables used in each regression block/step.

What You Have Learned

In this chapter you were introduced to logistic regression, which is a type of non-linear regression that is commonly used by social scientists. Logistic regression allows researchers to make predictions about dichotomous dependent variables. It relies on mathematically transforming a dichotomous dependent variable so that it ranges from negative infinity to positive infinity. As a result, the slope coefficients estimated by logistic regressions are the natural log of the odds that the outcome captured in the dependent variable will occur. Researchers usually transform these slope coefficients back into odds ratios in order to interpret logistic regression results. Researchers sometimes also present logistic regression results in the form of predicted probabilities. Understanding the basics of logistic regression gives you the conceptual background needed to understand other types of regression for categorical dependent variables, such as ordinal regression and multinomial regression. Although a full discussion of these more advanced regression techniques is beyond the scope of this book, you may encounter them in the academic literature or learn more about them if you take an advanced statistics course.

The research focus of this chapter was food security and food insecurity in Canada. Access to sufficient amounts of culturally appropriate food is a basic human right. In Canada, about 8 per cent of adults experience food insecurity, or go hungry. Not surprisingly, food insecurity is strongly related to income. People with incomes of less than $20,000 per year are more likely to go hungry than people with higher incomes. Younger adults have a higher probability of going hungry than older adults. Although food banks may provide some relief, they do little to alleviate the larger problem of food insecurity: many low-income households encounter barriers to accessing food banks, and if they do access them, they do not provide enough nutritional resources to consistently prevent people from going hungry.

Check Your Understanding

Check to see if you understand the key concepts in this chapter by answering the following questions:

1. When do researchers use logistic regressions instead of linear regressions?
2. Why is the dependent variable transformed in a logistic regression?
3. What are odds and how are they calculated?
4. What does it mean to find the “natural log” of a variable?
5. What are odds ratios, and how are they calculated? What does it mean when an odds ratio is less than 1, is exactly 1, or is greater than 1?
6. What does the “Exp(b)” column of an SPSS logistic regression show? What is the difference between how it is interpreted for slope coefficients and for the constant coefficient?
7. How are the standardized slope coefficients of the independent variables in a logistic regression calculated? What do they show?
8. What is Nagelkerke’s R^2? How is it interpreted?
9. Why is it important to identify whether researchers are reporting odds ratios or probabilities when they describe logistic regression results?

Practice What You Have Learned

Check to see if you can apply the key concepts in this chapter by answering the following questions. Keep *three* decimal places in all of the calculations in this chapter.

1. Calculate the odds associated with each of the following probabilities:
   a. 0.06
   b. 0.40
   c. 0.65
   d. 0.92
2. Calculate the natural log of each of the odds that you found in question 1.
3. Your local student union circulates a petition to establish a food bank on campus. Overall, 8 per cent of first-year students sign the petition, and among students in all other years, 16 per cent sign the petition.
   a. Calculate the odds that a first-year student will sign the petition.
   b. Calculate the odds that students in other years will sign the petition.
   c. Calculate the odds ratio that shows how the odds that a first-year student will sign the petition compare to the odds that students in other years will sign the petition.
4. Table 15.11 shows the results of a logistic regression. The dependent variable captures whether or not people formally volunteered for a group or organization in the past 12 months. (A "1" value indicates they volunteered, and "0" indicates they did not volunteer.) The independent variable is sex/gender, captured in a dummy variable. Explain what the odds ratio of the "Women" dummy variable shows.

Table 15.11 Results of a Logistic Regression with a Dummy Variable as an Independent Variable

Dependent variable: Formally volunteered for a group or organization in the past 12 months (n = 13,623)

	Unstandardized Coefficient	Odds Ratio
Women	0.108*	1.11
Constant	–0.311*	0.73
Nagelkerke R^2	*0.001*	

*Indicates that results are statistically significant at the p < 0.05 level.
Source: Author generated; Calculated using data from Statistics Canada, 2015.

5. Using the information in Table 15.11:
   a. Calculate the z-value for women and for men.
   b. Calculate the predicted probability of volunteering for women and for men.
6. Table 15.12 shows the results of a logistic regression predicting whether or not people formally volunteered for a group or organization in the past 12 months, using a ratio-level "Age" variable as the independent variable. Explain what the odds ratio of the "Age" variable shows.
7. Using the information in Table 15.12:
   a. Calculate the z-value for people at different ages, ranging from 20 to 80 (i.e., age 20, 30, 40, 50, 60, 70, and 80).

Table 15.12 Results of a Logistic Regression with a Ratio-Level Independent Variable

Dependent variable: Formally volunteered for a group or organization in the past 12 months (n = 13,623)

	Unstandardized Coefficient	Odds Ratio
Age (in years)	–0.012*	0.99
Constant	0.276*	1.32
Nagelkerke R^2	*0.015*	

*Indicates that results are statistically significant at the p < 0.05 level.
Source: Author generated; Calculated using data from Statistics Canada, 2015.

   b. Calculate the predicted probability of volunteering for people at each age.
   c. Either by hand or using a spreadsheet program, create a graph showing the predicted probability of volunteering for people at each age.
8. Table 15.13 shows the results of a logistic regression predicting whether or not people formally volunteered for a group or organization in the past 12 months, using both the ratio-level "Age" variable and the "Women" dummy variable as independent variables.
   a. Explain what the odds ratio of the "Age" variable shows. Be sure to pay attention to the idea of "controlling."
   b. Explain what the odds ratio of the "Women" dummy variable shows.

Table 15.13 Results of a Logistic Regression with Two Independent Variables (One Ratio-Level Variable and One Dummy Variable)

Dependent variable: Formally volunteered for a group or organization in the past 12 months (n = 13,623)

	Unstandardized Coefficient	Odds Ratio
Age (in years)	−0.012*	0.99
Women	0.124*	1.13
Constant	0.219*	1.24
Nagelkerke R^2	*0.016*	

*Indicates that results are statistically significant at the $p < 0.05$ level.
Source: Author generated; Calculated using data from Statistics Canada, 2015.

9. Using the information in Table 15.13, calculate the predicted probability of volunteering for each of the following people:
   a. An 18-year-old man
   b. A 45-year-old woman
   c. A 65-year-old man
10. Using the information in Table 15.13 and the same approach as in question 9, calculate the predicted probability of volunteering for men and women at different ages, ranging from 20 to 80 (i.e., age 20, 30, 40, 50, 60, 70, and 80). Either by hand or using a spreadsheet program, create a graph showing the predicted probability of volunteering for men and women at each age.
11. For the logistic regression shown in Table 15.13, the standard deviation of the dependent variable ("Formally volunteered for a group or organization in the past 12 months") is 0.496, the standard deviation of the "Age" variable is 18.273, and the standard deviation of the "Women" dummy variable is 0.500.
    a. Calculate the standardized slope coefficient of the "Age" variable.
    b. Calculate the standardized slope coefficient of the "Women" dummy variable.
    c. Determine which independent variable has the strongest relationship with the dependent variable, using the standardized slope coefficients.
12. Table 15.14 shows a nested logistic regression with three blocks, predicting whether or not people formally volunteered for a group or organization in the past 12 months. The first block uses the same independent variables as the regression shown in Table 15.13. Similar to the linear regression shown in Table 13.8, the second block adds variables capturing socio-economic characteristics, and the third block adds indicators of community engagement, including whether or not people donated money to a charitable organization in the past 12 months, and whether or not they participate in religious activities/services at least once a month.
    a. Explain what the odds ratio of the "Annual personal income" variable in the second model shows. Be sure to pay attention to the idea of "controlling."
    b. Explain what the odds ratios of three education dummy variables in the second model show.
    c. Describe the general pattern of the relationship between education and volunteering shown in the second model.
    d. Taking both education and income into account, describe how socio-economic status appears to be associated with volunteering. How might you explain this result?
13. Using the information in Table 15.14:
    a. Explain what the odds ratio of the "Donated money to a charitable organization in the past 12 months" dummy variable in the third model shows.
    b. Explain what the odds ratio of the "Participates in religious activities/services once a month or more often" dummy variable shows.
    c. Describe how the odds ratios of the other independent variables change between the second and the third model.
14. Using the information in Table 15.14, describe how the Nagelkerke R^2 changes between the three models. Explain what these changes show.
15. The logistic regression model in Table 15.15, excerpted from a *Journal of School Health* article, predicts whether Canadian students in grades 6 to 12 consume the recommended amount of fruits and vegetables (as determined by Canada's Food

Table 15.14 Results of a Nested Logistic Regression, with Three Blocks

Dependent variable: Formally volunteered for a group or organization in the past 12 months (n = 13,623)

	Odds Ratio		
	Model 1	Model 2	Model 3
Personal Characteristics			
Age (in years)	0.99*	0.99*	0.98*
Women	1.13*	1.21*	1.13*
Socio-economic Characteristics			
Annual personal income (in thousands of dollars)		1.00*	1.00*
Education (ref: high school only)			
Less than high school		1.22*	1.27*
Post-secondary diploma		1.24*	1.21*
University degree		1.94*	1.86*
Community Engagement Characteristics			
Donated money to a charitable organization in the past 12 months			2.09*
Participates in religious activities/services once a month or more often			2.10*
Constant	*1.24**	*0.84**	*0.49**
Nagelkerke R^2	*0.02*	*0.05*	*0.10*

*Indicates that results are statistically significant at the p < 0.05 level.
Source: Author generated; Calculated using data from Statistics Canada, 2015.

Guide; recommended amounts vary depending on age and gender). Overall, only about 10 per cent of students in grades 6 to 12 eat enough fruits and vegetables. All of the independent variables (predictors) in the model are dummy variables, with the reference group identified. School achievement and students' weekly spending allowance are also controlled for in this regression, but not displayed in this table.

a. Explain what the odds ratio of the "Boys" dummy variable shows. Be sure to pay attention to p-value and the 99 per cent confidence interval.

b. Explain what the odds ratio of each of the six grade dummy variables show. Be sure to pay attention to the p-values.

c. Describe the general pattern of the relationship between students' grade and whether or not they meet the fruit and vegetable recommendations.

16. Using the information in Table 15.15:

a. Explain what the odds ratio of the eight provinces dummy variables show. Be sure to pay attention to the p-values.

b. Explain what the odds ratio of the five ethnicity dummy variables show. Again, be sure to pay attention to the p-values.

Table 15.15 Logistic Regression Analysis of Variables Related to the Odds of Meeting Fruit and Vegetable Consumption Recommendations, Grades 6–12, Canada, 2012–2013 YSS

Predictors	Meet FV Recommendations (%)	Meeting FV Recommendations among All Students (Model 1: N = 36,455) OR Adjusted (99% CI)	p-Value
Sex			
Girls (ref)	10.7	1.0	
Boys	9.2	0.89 (0.77, 1.03)	0.0450
Grade			
6 (ref)	17.9	1.0	
7	17.1	0.92 (0.75, 1.14)	0.3188
8	10.9	0.51 (0.41, 0.65)	<0.000
9	5.6	0.23 (0.18, 0.31)	<0.000
10	6.3	0.25 (0.18, 0.33)	<0.000
11	6.6	0.22 (0.16, 0.30)	<0.000
12	6.8	0.22 (0.15, 0.31)	<0.000
Provinces			
Ontario (ref)	8.7	1.0	
Newfoundland	7.4	0.69 (0.54, 0.89)	0.0001
Prince Edward Island	8.3	0.70 (0.53, 0.91)	0.0005
Nova Scotia	9.4	0.72 (0.57, 0.91)	0.0003
New Brunswick	9.6	0.89 (0.71, 1.12)	0.2005
Quebec	10.2	0.83 (0.68, 1.02)	0.0186
Saskatchewan	9.4	0.83 (0.66, 1.05)	0.0416
Alberta	9.9	0.91 (0.73, 1.13)	0.2544
British Columbia	10.3	0.99 (0.81, 1.21)	0.9171
Ethnicity			
White (ref)	9.5	1.0	
Black	10.4	1.18 (0.85, 1.64)	0.1850
Asian	11.6	1.11 (0.88, 1.42)	0.2464
Aboriginal	10.2	1.25 (0.95, 1.65)	0.0402
Latin American	13.9	1.56 (0.97, 2.52)	0.0164
Other	11.6	1.33 (1.01, 1.78)	0.0098

CI, confidence interval; FV, fruit and vegetable; N, number; OR, odds ratio; YSS, Youth Smoking Survey.

Source: Excerpt from Minaker and Hammond 2016, 139.

Practice Using Statistical Software (IBM SPSS)

Answer these questions using IBM SPSS and the GSS27.sav or the GSS27_student.sav dataset available from the Student Resources area of the companion website for this book. Weight the data using the "Standardized person weight" [STD_WGHT] variable you created following the instructions in Chapter 5. Report two decimal places in your answers, unless fewer are printed by IBM SPSS. It is imperative that you save the dataset to keep any new variables that you create.

Note: The Binary Logistic procedure needed to answer these questions is only available in the Standard, Professional, and Premium editions of IBM SPSS; it is not available in the Base edition.

1. The variable "Victim of Discrimination – 5 years" [DISCRIM] shows people's answers to a sequence of questions asking people the following: "In the past five years, have you experienced discrimination or been treated unfairly by others in Canada because of your . . . sex, ethnicity or culture, race or colour, physical appearance, religion, sexual orientation, age, a disability, language, or some other reason?" Use the Recode into Different Variables tool to recode this variable into an "Experienced discrimination" [DISCRIM_RECODED] dummy variable to use as the dependent variable in a logistic regression. In the new variable, assign the value "1" to people who have been the victim of discrimination in the past five years (for any reason), and assign the value "0" to people who have not been the victim of discrimination in the past five years. The remaining values can be designated as system-missing in the new variable. Produce frequency distributions of the original variable "Victim of Discrimination – 5 years" [DISCRIM] and the new variable "Experienced discrimination" [DISCRIM_RECODED] and compare them to be sure that the recoding is correct.
2. Use the Binary Logistic Regression procedure to produce a regression of the independent variable "Visible minority" [IS_VISMIN] (you created this variable in question 2 of "Practice Using Statistical Software" in Chapter 13) on the dependent variable "Experienced discrimination" [DISCRIM_RECODED]. Explain what the odds ratio of the "Visible minority" dummy variable shows.
3. Use the options in the Binary Logistic Regression procedure to generate 95 per cent confidence intervals for the coefficients of the regression of the independent variable "Visible minority" [IS_VISMIN] on the dependent variable "Experienced discrimination" [DISCRIM_RECODED]. Explain what the confidence interval for the odds ratio of the "Visible minority" dummy variable shows.
4. Use the Binary Logistic Regression procedure to produce a regression of the independent variable "Age" [AGE] (you created this variable in question 1[a] of "Practice Using Statistical Software" in Chapter 13) on the dependent variable "Experienced discrimination" [DISCRIM_RECODED]. Explain what the odds ratio of the "Age" variable shows.
5. Use the Binary Logistic Regression procedure to produce a regression of the independent variables "Age" [AGE], "Visible minority" [IS_VISMIN], and "Women" [WOMEN] (you created this variable in question 3 of "Practice Using Statistical Software" in Chapter 12) on the dependent variable "Experienced discrimination" [DISCRIM_RECODED]. Explain what the odds ratios of the three independent variables show. Be sure to pay attention to the idea of "controlling."
6. Use the Save option in the Binary Logistic Regression procedure to save the predicted probabilities generated by the regression of the independent variables "Age" [AGE], Visible minority" [IS_VISMIN], and "Women" [WOMEN] on the dependent variable "Experienced discrimination" [DISCRIM_RECODED]. Use the Means procedure with the newly saved "Predicted probability" [PRE_1] variable to identify the lowest (minimum), highest (maximum), and average predicted probability of experiencing discrimination.
7. Use the Select Cases tool to select cases if the "Predicted probability" [PRE_1] is greater than or equal to "0" *and* the "Experienced discrimination" [DISCRIM_RECODED] variable is greater than or equal to "0". This ensures that only cases that are used in

the logistic regression model you produced in question 6 are used to calculate the statistics.

a. Find the standard deviation of the dependent variable, "Experienced discrimination" [DISCRIM_RECODED].
b. Find the standard deviation of the three independent variables: "Age" [AGE], Visible minority" [IS_VISMIN], and "Women" [WOMEN].
c. Use the standard deviations to calculate the standardized slope coefficients of each of the three independent variables. (For this question, keep three decimal places in your calculations.)
d. Determine which independent variable has the strongest relationship with the dependent variable, using the standardized slope coefficients.

After completing this question, use the Select Cases tool to return to using all of the cases.

8. Create three dummy variables to capture people's religious affiliation. Use the Recode into Different Variables tool to recode the variable "Religion of respondent - 7 categories" [RELIG7] into dummy variables as follows:

a. Create the new dummy variable "No religion" [NORELIGION] by assigning the old value "7" the new value "1", and assigning the old values "1" through "6" the new value "0". (The remaining values can be designated as system-missing in the new variable.)
b. Create the new dummy variable "Christian" [CHRISTIAN] by assigning the old value "2" the new value "1", and assigning the old value "1" and the old values "3" through "7" the new value "0". (The remaining values can be designated as system-missing in the new variable.)
c. Create the new dummy variable "Other religion" [OTHER_RELIG] by assigning the old values "1", "3", "4", "5", and "6" the new value "1", and assigning the old values "2" and "7" the new value "0". (The remaining values can be designated as system-missing in the new variable.)
d. Produce frequency distributions of the original variable "Religion of respondent - 7 categories" [RELIG7] and each of the three new dummy variables "No religion" [NORELIGION], "Christian" [CHRISTIAN], and "Other religion" [OTHER_RELIG], and compare them to be sure the recoding is correct.

9. Use the Binary Logistic Regression procedure to produce a nested regression, using "Experienced discrimination" [DISCRIM_RECODED] as the dependent variable. In the first block, use the same three independent variables as in the regression in questions 5 and 6: "Age" [AGE], Visible minority" [IS_VISMIN], and "Women" [WOMEN]. In the second block, add the variables "Christian" [CHRISTIAN] and "Other religions" [OTHER_RELIG]. ("No religion" [NORELIGION] is the reference group.)

a. Explain what the odds ratios of the two religion dummy variables in the second block show.
b. Compare the odds ratios of the "Age," "Visible minority" and "Women" variables in the first and second blocks. How does the magnitude of the odds ratio of each of these three variables change after religion is controlled for?
c. Describe how the Nagelkerke R^2 of the second model compares to that of the first model, and explain what this result shows.

Key Formulas

Odds of a probability

$$odds = \frac{\Pr(y_i = 1|x_i)}{1 - \Pr(y_i = 1|x_i)}$$

Odds of a probability (simplified)

$$odds = \frac{p_i}{1 - p_i}$$

Log odds	$log\ odds = ln\left(\frac{p_i}{1-p_i}\right)$
Logistic regression predictions (one independent variable)	$ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = a + bx$
Predicted probability	$\hat{p} = \frac{e^z}{1+e^z}$
Z-value (for predicted probability)	$z = a + b_1x_1 + b_2x_2 + b_3x_3 + \cdots$

References

Food Banks Canada. 2015. “Hunger Count.” Toronto. http://www.foodbankscanada.ca/FoodBanks/MediaLibrary/HungerCount/HungerCount2015_singles.pdf.

Health Canada. 2012. “Determining Food Security Status.” *Food and Nutrition*. July 25. http://www.hc-sc.gc.ca/fn-an/surveill/nutrition/commun/insecurit/status-situation-eng.php.

Loopstra, Rachel, and Valerie Tarasuk. 2012. “The Relationship between Food Banks and Household Food Insecurity among Low-Income Toronto Families.” *Canadian Public Policy* 38 (4): 497–514. doi:10.3138/CPP.38.4.497.

MacRae, Rod. 2012. “Food Policy for the Twenty-First Century.” In *Critical Perspectives in Food Studies*, edited by Mustafa Koç, Jennifer Sumner, and Anthony Winson, 310–23. Don Mills, ON: Oxford University Press.

Minaker, Leia, and David Hammond. 2016. “Low Frequency of Fruit and Vegetable Consumption among Canadian Youth: Findings from the 2012/2013 Youth Smoking Survey.” *Journal of School Health* 86 (2): 135–42. doi:10.1111/josh.12359.

Roshanafshar, Shirin, and Emma Hawkins. 2015. “Food Insecurity in Canada.” Catalogue no. 82–624X. Ottawa: Statistics Canada.

Statistics Canada. 2013. “Canadian Community Health Survey (CCHS) Annual Component User Guide 2012 and 2011–2012 Microdata Files.” Ottawa: Statistics Canada.

———. 2014. Canadian Community Health Survey, 2012: Annual Component. *Public Use Microdata File*. Ottawa, ON: Statistics Canada.

———. 2015. General Social Survey, 2013: Cycle 27, Giving, Volunteering and Participating. *Public Use Microdata File*. Ottawa, ON: Statistics Canada.

Tarasuk, Valerie, Naomi Dachner, and Rachel Loopstra. 2014. “Food Banks, Welfare, and Food Insecurity in Canada.” Edited by Professor Martin Caraher and Dr Alessio Cavic. *British Food Journal* 116 (9): 1405–17. doi:10.1108/BFJ-02–2014–0077.

UN Food and Agriculture Organization (UNFAO). 2001. *The State of Food Insecurity in the World 2001*. Rome: Food and Agriculture Organization of the United Nations. http://www.fao.org/docrep/003/y1500e/y1500e00.htm.

Vozoris, Nicholas T., and Valerie S. Tarasuk. 2003. “Household Food Insufficiency Is Associated with Poorer Health.” *Journal of Nutrition* 133 (1): 120–26.