

# UBY/UTUBYL ILL – Lending

Call #: QA 278.2 .L586 2023

Location: 2ND

Journal Title: Categorical data analysis and multilevel modeling using R /

Volume: Issue:

Month/Year: 2023 Pages: 95

Article Author: Xing Liu

Article Title: Chapter 3. Logistic Regression for Binary Data

Ref. Number / Item Barcode:



ILL Number: -21738744



Lending String:

Borrower: RAPID:PUL

NEW: Interlibrary Services,  
Firestone

Odyssey: 128.112.202.139

Email: [ilsborr@princeton.edu](mailto:ilsborr@princeton.edu)  
12/8/2023 12:55 PM



TN: 2632585

Pages: 95 RAPID:PUL

Exempt

# BYU

BRIGHAM YOUNG UNIVERSITY

# LIBRARY

## ODYSSEY PROBLEM REPORT

If you have experienced a problem in the delivery of the requested item, please contact us within 5 Business Days. Please include the following information:

ILL #: \_\_\_\_\_

Your OCLC Symbol: \_\_\_\_\_

Date of Receipt: \_\_\_\_\_

Please Specify if:

- Pages were missing pp. \_\_\_\_\_ to \_\_\_\_\_
- Edges were cut off pp. \_\_\_\_\_ to \_\_\_\_\_
- Illegible copy, resend entire copy.
- Wrong article sent.
- Other (explain)

## FOR ODYSSEY RECEIPT PROBLEMS:

Phone: 801-422-6844

Email: [library\\_request@byu.edu](mailto:library_request@byu.edu)

## NOTICE:

This material may be  
protected by copyright law  
TITLE 17 U.S. Code

# 3

## LOGISTIC REGRESSION FOR BINARY DATA

### OBJECTIVES OF THIS CHAPTER

This chapter introduces logistic regression models for binary data. It introduces concepts of odds, odds ratio, and goodness-of-fit statistics of the model; describes how to test the significance of predictors; and shows how to interpret parameter estimates. Following the description of data, two logistic regression models with the `glm()` function are illustrated with step-by-step instructions. R commands and output are explained in detail. In addition, probit regression with R is also introduced. The focus of this chapter is on fitting binary logistic regression models using R, as well as on interpreting and presenting the results. After reading this chapter, you should be able to:

- Determine when a logistic regression model is used.
- Conduct logistic regression using R.
- Interpret the output.
- Interpret the model in terms of odds ratios.
- Interpret the marginal effects
- Compute and plot the predicted probabilities.
- Compare models using the likelihood ratio test and other fit statistics.
- Present the results in publication-quality tables.
- Write the results for publication.



## 3.1 LOGISTIC REGRESSION MODELS: AN INTRODUCTION

---

In multiple linear regression, a set of independent variables is used to predict a continuous outcome variable. In this chapter, logistic regression models are introduced for the analysis of dichotomous response variables with two categories.

Research examples include binary response variables such as life satisfaction (Hoffmann, 2016), having a fracture or not (Hosmer et al., 2013), gun ownership (Kaufman, 2019), computer use (Menard, 2010), and skipping school or not (Smithson & Merkle, 2014). These binary response variables have two categories: having an event or not having an event. They are normally coded as 1 for having an event and 0 for not having an event.

We can use multiple linear regression as a preliminary method to estimate a binary outcome variable. The method is referred to as the linear probability model. Due to its simplicity in computation and easy interpretation as with the linear regression model, the linear probability model is still popular in economics. However, this model has several drawbacks because using multiple regression with the ordinary least squares method to estimate a binary response variable violates the assumptions of normality, homoscedasticity, and the linear relationship between the dependent variable and independent variables. First, in multiple linear regression, we assume that the dependent variable has a normal distribution. When the dependent variable is binary, it has a Bernoulli distribution, which is not a normal distribution. Second, the homoscedasticity assumption suggests that the error variances of the dependent variable are the same across each value of the independent variables. This assumption is violated because the error variances depend on the values of the predictor variables and vary across them, which is referred to as heteroscedasticity. Heteroscedasticity may lead to biased standard errors of the estimates and invalid statistical inference. Third, multiple linear regression also assumes that the relationship between the dependent variable and independent variables is linear. When the dependent variable is dichotomous, the relationship between this variable and an independent variable is a nonlinear, S-shaped curve, which can be easily seen if we draw a scatterplot showing the relationship between the dichotomous dependent variable and the independent variable. In addition, we may see the predicted values of the dependent variable are out of the range of 0 and 1. In other words, the estimated values of the outcome variable may be negative (i.e., smaller than 0) or larger than 1. These values are nonsensical since they are the predicted probabilities from the independent variables. Therefore, we should use the linear probability model with caution. If we do use this model, the robust standard error approach should be used to adjust the standard errors to obtain unbiased standard errors of the estimates.

### 3.1.1 Why Do We Need a Logistic Transformation?

As explained previously, if we use linear regression to estimate binary data, heteroscedasticity (i.e., unequal error variance) is present in the model and the predicted value may have values less than 0 or larger than 1. To estimate the probability of success for having an event, a logistic transformation  $\text{logit}(p)$  or a probit transformation  $\text{probit}(p)$

can be used. A regression model with the logistic transformation is called the logistic regression model, whereas a model with the probit transformation is called the probit regression model. Both models produce similar results and use of either model is down to the choice of researchers. This chapter introduces the binary logistic regression models first and then the probit regression models.

The form of the simple logistic regression model can be expressed as follows:

$$\text{logit}(p) = \alpha + \beta X$$

where  $p$  is the probability when the outcome variable equals 1,  $P(Y = 1)$ ;  $\text{logit}(p)$  is the logistic transformation of the probability of success or of an event occurrence; and on the right side,  $\alpha$  is the intercept and  $\beta$  is the logit regression coefficient.

This equation looks like that for simple linear regression. The noticeable difference is the logistic transformation on the left side of logistic regression. Instead of directly estimating the dependent variable, we estimate the logistic transformation (i.e., logit) of the probability of a success, which is also known as the logarithm of the odds or “log odds.” Odds are the ratio of the probability of success to the probability of failure. The transformation between probabilities and odds will be introduced in detail in the next section. Therefore, in simple logistic regression, we estimate the relationship between an independent variable and the binary outcome variable on a scale of the logit or log odds. In other words, the relationship between the predictor variable  $X$  and the logit of the outcome (i.e., the logit transformation of the probability when the outcome variable  $Y = 1$ ) is linear. The logit of the outcome  $\text{logit}(p)$  can be easily transformed back to the probability of the outcome  $P(Y = 1)$  since the logit is the natural logarithm of the odds or log odds.

Since  $\text{logit}(p)$  is  $\ln(\text{odds})$ , which is expressed as  $\ln \frac{p}{1-p}$ , the form of the simple logistic regression can also be rewritten as:

$$\ln \frac{p}{1-p} = \alpha + \beta X \quad (3.1)$$

where  $\ln$  is the natural logarithm. For simplicity, we read “ $\ln(\text{odds})$ ” log odds or log of the odds.

When the probability ( $p$ ) varies from 0 to 1, the log odds or logit will vary from negative infinity to positive infinity.

In multiple logistic regression, we have more than one predictor variable. The following is the form of the multiple logistic regression model:

$$\ln \frac{p(x)}{1-p(x)} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p \quad (3.2)$$

where  $X_1, X_2, \dots, X_p$  are the predictor variables and  $\beta_1, \beta_2, \dots, \beta_p$  are the logit coefficients of these predictors. This equation can be also expressed as:

$$\text{logit}[p(x)] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p \quad (3.3)$$

### 3.1.2 Probabilities, Odds, and Odds Ratios

Just as the mean and standard deviation are the cornerstone for the descriptive statistics and linear models, probabilities, odds, and odds ratios are the key concepts for logistic regression models. Probability is the chance of success or of having an event occur. It can be expressed as a proportion. It equals the frequency of success or of having an event divided by the total number of observations or events. Probabilities range from 0 to 1. When probability is 0, it means that the chance of success is 0 or that there is no chance an event will occur. When probability is 1, it means that the chance of success is certain.

If  $p$  is the probability of success or having an event, then the probability of failure or of not having an event is  $1 - p$  since these two probabilities are complementary. The sum of all possible probabilities equals 1.

The odds are the ratio of two probabilities, the probability of success or of having an event ( $p$ ) to the probability of failure or of not having an event ( $1 - p$ ). Since  $p = P(Y = 1)$  and  $1 - p = P(Y = 0)$ , the odds can be expressed as follows.

$$\text{Odds} = \frac{P(Y = 1)}{P(Y = 0)} = \frac{p}{1 - p}$$

Since the probability  $p$  varies from 0 to 1, the odds vary from 0 to positive infinity. With the increase of the probability of success or of having an event, the odds also get larger. Let us see three examples:

$$\text{When } p = 0.1, \text{ odds} = 0.1/(1 - 0.1) = 0.11.$$

$$\text{When } p = 0.5, \text{ odds} = 0.5/(1 - 0.5) = 1.00.$$

$$\text{When } p = 0.9, \text{ odds} = 0.9/(1 - 0.9) = 9.00.$$

The preceding examples focus on the odds of having an event. What are the odds of not having that event? When the odds of having an event = 99, the odds of not having an event =  $1/99 = 0.01$  since they are just the multiplicative odds or the reciprocal of the odds of having an event. This is an important property of odds. It is useful when we compare different categories for the odds in various ordinal logistic regression models in the following chapters.

The examples show us how to compute odds if we know the probability of success or of having an event. But if we know the odds of success or of having an event, can we compute the probability of success or of having an event? This backward transformation is also easy. Since  $\text{odds} = \frac{p}{1 - p}$ , with a simple transformation, we get  $p = \frac{\text{odds}}{1 + \text{odds}}$ . In other words, the probability of success or of having an event equals the odds divided by  $(1 + \text{odds})$ .

For example, the command `table(healthy, wrkfull)` produces a two-way cross-tabulation table of two binary variables, the health status and working full-time.

The health status has two categories with 1 = good health and 0 = not good health; the working full-time is coded as full-time and not.

```
> tab <- table(healthy, wrkfull)
> tab
```

	wrkfull	
healthy	0	1
0	363	182
1	635	693

```
> ftable(tab)
```

	wrkfull	
healthy	0	1
0	363	182
1	635	693

```
> addmargins(tab)
```

	wrkfull		
healthy	0	1	Sum
0	363	182	545
1	635	693	1328
Sum	998	875	1873

Since the probability of success or of having an event is a ratio of the frequency of success or of having an event to the total observations, the probability of having good health  $P(\text{healthy} = 1) = 1,328/1,873 = .709$ . The probability of not having good health  $P(\text{healthy} = 0) = 545/1,873 = .291$ . Therefore, overall, the odds of having good health =  $P(\text{healthy} = 1)/P(\text{healthy} = 0) = .709/.291 = 2.436$ .

Next, let us compute the odds of having good health for not working full-time.

To compute the probability of having good health for people who are currently not working full-time, we look at the ratio of the frequency of people with good health who are not working full-time to the total frequency of people who are not working full-time:  $P(\text{healthy} = 1 \mid \text{not full-time work}) = 635/998 = .636$ .

The probability of having poor health for people who do not work full-time  $P(\text{healthy} = 0 \mid \text{not full-time work}) = 363/998 = .364$ .

For people who do not work full-time, the odds of having good health =  $P(\text{healthy} = 1 \mid \text{not full-time work})/P(\text{healthy} = 0 \mid \text{not full-time work}) = .636/.364 = 1.747$ .

For full-time work, the odds of having good health =  $P(\text{healthy} = 1 \mid \text{full-time work})/P(\text{healthy} = 0 \mid \text{full-time work})$ . Since  $P(\text{healthy} = 1 \mid \text{full-time work}) = 693/875 = .792$  and  $P(\text{healthy} = 0 \mid \text{full-time work}) = 182/875 = .208$ , the odds of having good health for the people who work full-time =  $.792/(1 - .792) = 3.808$ .

### 3.1.3 Transformation Among Probabilities, Odds, and Log Odds in Logistic Regression

#### Forward Transformation

If we know the probability of success or of having an event, the odds of success are just the probability of success divided by the probability of failure. By taking the natural logarithm of odds, we get log odds or logit. Table 3.1 presents the forward transformation from probabilities to odds to log odds.

**TABLE 3.1** ◉ Forward Transformation From Probabilities to Odds to Log Odds

Transformation	Probabilities	Odds	Logit or Log Odds
Forward	$p$	$\frac{p}{1-p}$	$\ln \frac{p}{1-p}$

#### Backward Transformation

If we know the logit, since it is the natural logarithm of the odds or log odds, then the odds are the antilogarithm. In other words, the odds of success can be obtained by taking the exponential of the logit. To transform the odds back to the probability of success, we divide the odds by 1 plus the odds. The backward transformation is displayed in Table 3.2.

**TABLE 3.2** ◉ Backward Transformation From Log Odds to Odds to Probabilities

Transformation	Logit or Log Odds	Odds	Probabilities
Backward	$\text{Logit}(p)$ or $\ln \frac{p}{1-p}$	$\text{Odds} = \exp(\text{logit})$	$p = \frac{\text{odds}}{1 + \text{odds}}$

#### Odds Ratio

What is an odds ratio (OR)? It is just a ratio of two odds. Since the odds of having good health for the people who work full-time are 3.808 and the odds of having good health for not working full-time are 1.747, the ratio of the odds for working full-time to the odds for not working full-time =  $3.808/1.747 = 2.180$ . In other words, the odds of having good health for working full-time are 2.180 times the odds for not working full-time.

When  $OR > 1$ , the odds of success or of having an event for one group are larger than the odds for the other group. For example,  $OR = 2$  indicates the odds of success for one group are two times the odds for the other group.

When  $OR < 1$ , the odds in one group are less than the odds in the other group. For example,  $OR = 0.5$  indicates that the odds for one group are 0.5 times the odds for the other group. In other words, the odds for the second group are twice the odds for the first group. When  $OR$  is less than 1, we can take the inverse of it and make it more interpretable.

When  $OR = 1$ , the odds for one group are the same as the ones for the other group.

### 3.1.4 Bernoulli Distributions, the Likelihood Function, and Maximum Likelihood Estimation

#### Bernoulli Distributions

When a discrete random variable is a binary outcome variable with 1 for having an event or success and 0 for not having an event or failure, it follows a Bernoulli distribution. When  $Y = 1$ , the probability of success or of having an event,  $p = P(Y = 1)$ ; when  $Y = 0$ , the probability of failure or of not having an event,  $1 - p = P(Y = 0)$ . The probability function, also called the probability mass function (PMF), for Bernoulli distributions is expressed as:

$$P(Y = y) = p^y(1 - p)^{1-y} \quad (3.4)$$

where  $y$  can be either 1 or 0.

For each observation, we add subscripts to the equation as follows.

$$P(Y = y_i) = p_i^{y_i}(1 - p_i)^{1-y_i} \quad (3.5)$$

where  $y_i$  can be either 1 or 0.

With the Bernoulli probability function or PMF, we can obtain the probability at each value of the discrete random variable. For all the observations in the sample, you can obtain the joint probability function by multiplying the probability function of each independent observation as follows.

$$P(Y = y_1, y_2, \dots, y_n) = \prod_{i=1}^n p_i^{y_i}(1 - p_i)^{1-y_i} \quad (3.6)$$

where  $\prod$  is the product term for the joint probability.

#### The Likelihood Function

By reparametrizing the function above, we obtain the likelihood function.

$$L(p; y) = \prod_{i=1}^n p_i^{y_i}(1 - p_i)^{1-y_i} \quad (3.7)$$

where  $L(p; y)$  is the likelihood function, which is the function of the unknown parameter  $p$  given the observed data  $y$ . The likelihood function is not the same as the

probability function, although both equations on the right-hand side look the same. In the likelihood function, the parameter is unknown and needs to be estimated, whereas in the probability function we use the known parameter to obtain probabilities at different values of the discrete random variable. The likelihood function is often expressed as the form of the log likelihood function by taking the natural log of the function in Equation 3.7, so we can work on addition rather than multiplication of the function of individual observations.

The log likelihood function for the Bernoulli distribution can be expressed as:

$$\ln[L(p; y)] = \sum_{i=1}^n \{y_i \ln p_i + (1 - y_i) \ln(1 - p_i)\} \quad (3.8)$$

Since  $(1 - y_i) \ln(1 - p_i) = \ln(1 - p_i) - y_i \ln(1 - p_i)$ , the equation above can be rewritten as:

$$\ln[L(p; y)] = \sum_{i=1}^n \{y_i \ln p_i + \ln(1 - p_i) - y_i \ln(1 - p_i)\} = \sum_{i=1}^n \left\{ y_i \ln \frac{p_i}{1 - p_i} + \ln(1 - p_i) \right\} \quad (3.8b)$$

If we define  $l(p; y) = \ln[L(p; y)]$ , then the log likelihood function is expressed as:

$$l(p; y) = \sum_{i=1}^n \left\{ y_i \ln \frac{p_i}{1 - p_i} + \ln(1 - p_i) \right\} \quad (3.9)$$

## Maximum Likelihood Estimation

We use maximum likelihood estimation to estimate the unknown parameter by maximizing the likelihood in the function above. In other words, we look for the value of the parameter  $p$  which makes the log likelihood value  $l$  the largest. If we find that estimated value of the parameter  $p$ , then that value is the maximum likelihood estimate that maximizes the likelihood of observing the data.

Let us see an example of the sample data with 10 Bernoulli trials. The set of the sample data includes the value of each observation  $y_i$ , which is either 1 or 0

$$\{1, 0, 0, 1, 0, 1, 0, 1, 1, 1\}$$

The likelihood function is:

$$L(p; y) = p(1 - p)(1 - p)p(1 - p)p(1 - p)ppp = p^6(1 - p)^4$$

The log likelihood function is:

$$l(p; y) = \ln[p^6(1 - p)^4] = 6 \ln p + 4 \ln(1 - p)$$

To use maximum likelihood estimation to estimate the unknown parameter  $p$ , we take the derivative of the log likelihood function  $l(p; y)$  with respect to  $p$ , set it to 0, and find the solution to the equation.

$$\frac{d}{dp} l = \frac{d}{dp} [6 \ln p + 4 \ln(1 - p)] = \frac{6}{p} - \frac{4}{1 - p} = 0$$

Therefore, we obtain the estimated  $p$  or  $\hat{p} = \frac{6}{10} = \frac{3}{5}$ .

The  $\hat{p}$  value is the maximum likelihood estimate of the parameter  $p$ . With  $\hat{p} = \frac{3}{5}$ , the log likelihood function  $l(p; y)$  has its maximum value. In other words, among all the possible values for the parameter  $p$ , the estimated  $p = \frac{3}{5}$  is the maximum likelihood estimate which makes the sample data with 10 Bernoulli trials above to be the most likely.

Similarly, in simple logistic regression, we use maximum likelihood estimation to estimate the parameter, the logit coefficient  $\beta$ . We take the first derivative of the log likelihood function  $l(\beta; y, x)$  with respect to the coefficient  $\beta$ , set it to 0, and solve. To estimate the standard error of the coefficient, we take the second derivative of the log likelihood function with respect to the coefficient to obtain the Hessian matrix. The variance–covariance matrix of the coefficient is just the negative inverse Hessian matrix. Recall that in simple logistic regression, we estimate the relationship between an independent variable and the binary outcome variable on a scale of the logit or log odds and the probability of the outcome  $P(Y = 1)$  is conditional on the values of the predictor variable. Maximum likelihood estimation of the coefficient is far more complex than the example demonstrated above. Technical detail of maximum likelihood estimation is beyond the scope of this book; interested readers should refer to Hilbe (2009) for more detail.

## The Binomial Distribution

The Bernoulli distribution is a special case of the binomial distribution with one trial for each individual. When we count the number of successes for the total trials, the discrete random variable follows a binomial distribution. With  $k$  successes in  $n$  trials with the success probability  $p$ , the binomial distribution is expressed as:

$$P(Y = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad (3.10)$$

where  $\binom{n}{k}$  is the binomial coefficient and is read as “ $n$  choose  $k$ .” It gives us the number of ways we have  $k$  successes in  $n$  trials.

$$\binom{n}{k} = \frac{n!}{(n - k)!k!}$$

where  $n!$  is  $n$  factorial or the factorial of  $n$ .  $n! = n*(n - 1) \dots 2*1$ .



The log likelihood function for the binomial distribution is expressed as:

$$l(p; y; n) = \sum_{i=1}^n \left\{ y_i \ln \frac{p_i}{1-p_i} + n_i \ln(1-p_i) + \ln \binom{n}{k} \right\} \quad (3.11)$$

Since  $\ln \binom{n}{k}$  is a constant, when  $n = 1$ , the log likelihood function for the binomial distribution is the same as the log likelihood function for the Bernoulli distribution; so is the estimate by maximum likelihood estimation.

### 3.1.5 Goodness-of-Fit Statistics

To assess whether a model fits the data well, we normally look at several measures of fit statistics rather than at a single measure. The deviance, log likelihood ratio test, pseudo  $R^2$ , AIC, and BIC statistics are introduced next.

#### Deviance

Deviance is one of the goodness-of-fit statistics used in the logistic regression. It compares the currently fitted model and the saturated model. It is defined as  $-2$  times the difference in log likelihood between these two models, which can be expressed as  $-2(\log \text{ likelihood of the current model} - \log \text{ likelihood of the saturated model})$ . The saturated model is the model that fits the data perfectly.

In binary logistic regression, the saturated model estimates one parameter for each observation so the number of parameters equals the number of sample size. Its likelihood of fitting the data is 1, so the log likelihood of the saturated model equals  $\ln 1$ , which is 0. The expression of deviance can be simplified to  $-2 \log \text{ likelihood of the current model}$ . It is often abbreviated as  $-2LL$ .

Deviance shows how well a model fits the data compared with a saturated model that fits the data perfectly. If the discrepancy in log likelihood between the fitted model and the saturated model is small, then this model has a good fit. On the contrary, if the deviance is large, then the fitted model has a poor fit. Therefore, smaller deviance means a better fit. In a linear regression model, we minimize the error variance and would like to see the sum of squared residuals as small as possible. Similarly, in a logistic regression model, we would like to minimize the deviance and would like to see the discrepancy in log likelihood between the fitted model and the saturated model as small as possible.

#### Model Comparisons Using the Deviance Difference or Likelihood Ratio Test

Deviance is often used to compare nested models. Models are nested when more constraints can be put on parameters in one model than in the other. One model is called the reduced model, which contains less parameters, and the other is called the full model, within which the reduced model is nested but has more parameters. In the logistic regression, the reduced model has fewer variables than the full model, and the former model is a subset of the latter. The difference in deviance between nested

models follows a chi-square distribution. The degrees of freedom of the distribution equal the difference in the number of parameters between these two models. The difference in deviance is often expressed as  $G = \text{Deviance for the reduced model} - \text{Deviance for the full model}$  or as follows:  $G = D_{\text{Reduced}} - D_{\text{Full}}$ .

This test is also known as the likelihood ratio test since the difference in deviance is the difference in  $-2LL$ , which can be expressed as a ratio of likelihood in logarithm.

$$G = -2 \ln \frac{\text{Likelihood}_{\text{Reduced}}}{\text{Likelihood}_{\text{Full}}} = -2LL_{\text{Reduced}} - (-2LL_{\text{Full}}) = D_{\text{Reduced}} - D_{\text{Full}} \quad (3.12)$$

where  $D_{\text{Reduced}} = -2LL_{\text{Reduced}}$  and  $D_{\text{Full}} = -2LL_{\text{Full}}$ .

In simple logistic regression with only one independent variable, the likelihood ratio test compares the deviance between the null model with only the intercept ( $D_0$ ) and the model with one independent variable ( $D_m$ ). The likelihood ratio test,  $G = -2LL_0 - (-2LL_m) = D_0 - D_m$ . If the likelihood ratio chi-square test is significant, then we reject the null hypothesis and conclude that the model with one independent variable fits the data better than the model with only the intercept (null model).

## Predictor Selections Using the Likelihood Ratio Test

The likelihood ratio chi-square test can be used to test whether a predictor variable contributes to the model by comparing the models with and without the variable. A significant test means that the added variable contributes to the model. It is also useful for model developing. We can build a series of nested models from a simple model with one predictor to more complex models with multiple predictors. The likelihood ratio chi-square test can be used to decide which model fits the data better and whether the predictor variables should be kept in or removed from the model.

## Pseudo $R^2$

In linear regression models, the coefficient of determination,  $R^2$ , is the index for the model fit. It is a ratio of the variance explained by the model to the total variance. It indicates how much variance in the dependent variable is accounted for by an independent variable or a set of independent variables. Analogous to  $R^2$  in the linear regression, several pseudo  $R^2$  are used in logistic regression (Hardin & Hilbe, 2018; Long & Freese, 2014; Menard, 2010). However, the interpretation of these pseudo  $R^2$  in logistic regression is different from the interpretation of  $R^2$  in the linear regression. In addition, there is no consensus as to which of the pseudo  $R^2$  measures is the best. Table 3.3 displays three major pseudo  $R^2$  measures and their formulas.

### 1. The Likelihood Ratio $R^2$

The likelihood ratio  $R^2$ , written as  $R_L^2$ , is also known as McFadden's  $R^2$ , or the McFadden  $R^2$ . It is the reduction in deviance from the fitted model ( $D_m$ ) to the null model that only contains the intercept ( $D_0$ ). It is expressed as:

$$R_L^2 = 1 - \frac{D_m}{D_0} = 1 - \frac{-2LL_m}{-2LL_0} \tag{3.13}$$

when model deviance equals  $-2$  times log likelihood ( $D = -2LL$ ), the likelihood ratio  $R^2$  can also be the reduction in log likelihood between the fitted model and the null model. It is expressed as:

$$R_L^2 = 1 - \frac{-2LL_m}{-2LL_0} = 1 - \frac{LL_m}{LL_0} \tag{3.14}$$

2. Cox and Snell's  $R^2$

Cox and Snell's  $R^2$  or the Cox and Snell  $R^2$ , written as  $R_{ML}^2$ , is also known as the maximum likelihood  $R^2$ . It is based on the likelihood function of the fitted model ( $L_m$ ) and the null model, which only contains the intercept ( $L_0$ ). It is expressed as:

$$R_{ML}^2 \left( \frac{L_0}{L_m} \right)^{2/n} \tag{3.15}$$

where  $n$  is the total number of observations.

3. Nagelkerke's  $R^2$

This is also called Cragg and Uhler's  $R^2$ . It is an adjustment to the Cox and Snell  $R^2$  by dividing the maximum value of the Cox and Snell  $R^2$ . It is expressed as:

$$R_N^2 = R_{ML}^2 / \text{maximum } R_{ML}^2 \tag{3.16}$$

4. Other Pseudo  $R^2$  Measures

Other pseudo  $R^2$  measures include the adjusted McFadden's  $R^2$ , McKelvey and Zavoina's  $R^2$ , Efron's  $R^2$ , Tjur's  $R^2$ , and the Count and adjusted Count  $R^2$ . Refer to Hardin and Hilbe (2018) and Long and Freese (2014) for detailed introductions to pseudo  $R^2$  measures.

**TABLE 3.3** Three Major Pseudo  $R^2$  Measures

Pseudo $R^2$	Formulas
Likelihood ratio $R^2$ (McFadden's $R^2$ )	$R_L^2 = 1 - \frac{D_m}{D_0} = 1 - \frac{-2LL_m}{-2LL_0}$
Cox and Snell's $R^2$ (maximum likelihood $R^2$ )	$R_{ML}^2 \left( \frac{L_0}{L_m} \right)^{2/n}$
Nagelkerke's $R^2$ (Cragg and Uhler's $R^2$ )	$R_N^2 = R_{ML}^2 / \text{maximum } R_{ML}^2 = \left\{ 1 - \left( \frac{L_0}{L_m} \right) \right\}^{2/n} / (1 - L_0)^{2/n}$

## Information Criteria Indices: AIC and BIC

The likelihood ratio tests are used for comparisons between nested models. But how do you compare models if they are not nested (e.g., if two models contain different sets of independent variables or if one model has more missing values so the sample size is different between the two models)? To compare non-nested models, the Akaike information criterion (AIC) (Akaike, 1974) and the Bayesian information criterion (BIC) (Schwarz, 1978) statistics are commonly used. Both statistics are based on the deviance statistics and can be seen as an adjustment to the deviance. Just like the likelihood ratio test, they can also be used to compare nested models. Table 3.4 presents formulas for the AIC and BIC statistics.

The AIC adjusts or penalizes the deviance by the number of predictors:

$$\text{AIC} = -2(\text{LL}_m - k) = D_m + 2k \quad (3.17)$$

where  $k$  is the number of parameters (the number of independent variables plus the intercept) and  $-2\text{LL}_m$  or  $D_m$  is the deviance of the fitted model.

Remember that we would like to see smaller deviance; this is still true for the AIC statistic. A smaller AIC means a better fit of the model.

The BIC adjusts the deviance by its degrees of freedom and the sample size:

$$\text{BIC} = -2\text{LL}_m + \ln(n) \times k = D_m + \ln(n) \times k \quad (3.18)$$

where  $k$  is the number of parameters (the number of independent variables plus the intercept),  $n$  is the number of the sample size, and  $D_m$  is the deviance of the fitted model. Similar to the AIC statistic, we also prefer a smaller BIC statistic.

The AIC and BIC can be used to compare both nested and non-nested models. When comparing nested models, the BIC often leads to the selection of more parsimonious models, which are simpler models with few parameters. The BIC statistic has a larger penalty term in the form than that in the AIC statistic, so if the focus is to choose a more parsimonious model, the BIC is normally preferred.

**TABLE 3.4** Information Criteria Indices: AIC and BIC Statistics

Information Criteria Indices	Formulas
AIC	$-2(\text{LL}_m - k) = D_m + 2k$
BIC	$-2\text{LL}_m + \ln(n) \times k = D_m + \ln(n) \times k$

### 3.1.6 Testing Significance of Predictors

To test the statistical significance of each predictor, the Wald test is used. It is computed as a ratio of the parameter estimate for each predictor variable in the model to its corresponding standard error.

$$\text{Wald } z = \frac{\hat{\beta}}{\text{SE}(\hat{\beta})} \quad (3.19)$$

where  $\hat{\beta}$  is the estimated logit coefficient and  $\text{SE}(\hat{\beta})$  is the standard error. Statistical software packages report either the univariate Wald  $z$  statistic, which follows a standard normal distribution, or a Wald chi-square test statistic,  $z^2$ , which follows a chi-square distribution. The null hypothesis is that the coefficient of each predictor variable is 0. Rejection of the null hypothesis indicates that the effect of a particular predictor variable is significant.

$$\text{Confidence intervals for the parameter } \beta = \hat{\beta} \pm z * \text{SE}(\hat{\beta})$$

where  $\hat{\beta}$  is the estimated logit coefficient for a predictor,  $z$  is the  $z$ -score from a normal distribution for the chosen confidence interval, and  $\text{SE}(\hat{\beta})$  is the standard error. For the 95% confidence intervals,  $z = 1.96$ . Therefore, 95% confidence intervals for the parameter  $\beta = \hat{\beta} \pm 1.96 * \text{SE}(\hat{\beta})$ .

Another way to test the significance of predictors is the likelihood ratio test, which compares  $-2\text{LL}$  of different models. With only one additional predictor added to the existing model, the difference in deviance between the nested models has a chi-square distribution with one degree of freedom (i.e., the degree of freedom equals the difference between the number of predictors). A significant likelihood ratio test means that the variable is a significant predictor in the model. For a univariate test of a single predictor, the results of the Wald test and the likelihood ratio test are equivalent. The likelihood ratio test can also be used to test the contribution of two or more variables when they are added to the current model. A significant likelihood ratio test for the nested models means that a set of variables overall makes a significant contribution to the model.

### 3.1.7 Interpretation of Model Parameter Estimates in Logistic Regression

#### Probabilities, Odds, and Odds Ratios in Logistic Regression

In Section 3.1.3, the transformation from probabilities to odds and to odds ratios was introduced. Let us take a look at their transformation in logistic regression. In the simple logistic regression,  $\ln \frac{p}{1-p} = \alpha + \beta X$ , where  $p$  is the probability when  $Y = 1$ . The estimated coefficient is the logit coefficient, which is the coefficient on the scale of logit or log odds. In the simple logistic regression, the outcome variable is dichotomous with values of 1 and 0. We estimate the relationship between the predictor variable and the logit function of the probability that  $Y = 1$ , the log odds.

Exponentiating both sides of the equation, we get the odds of success or of having an event:

$$\text{Odds}(Y = 1) = \exp(\alpha + \beta X)$$

If the independent variable  $X$  is a categorical variable with the values of 0 and 1, what are the odds of success or of having an event?

When  $X = 0$ , odds ( $Y = 1$ ) =  $\exp(\alpha)$ , which is the exponentiated intercept.

When  $X = 1$ , odds ( $Y = 1$ ) =  $\exp(\alpha + \beta)$ , which is the exponentiated sum of intercept and logit coefficient.

The odds ratio of the group 1 ( $X = 1$ ) to the group 2 ( $X = 0$ ):

$$\text{OR} = \frac{\exp(\alpha + \beta)}{\exp(\alpha)} = \frac{\exp(\alpha) \times \exp(\beta)}{\exp(\alpha)} = \exp(\beta) \quad (3.20)$$

This is the simple case for a one-unit increase in an independent variable (e.g., from 0 to 1 in the previous example); the change in the odds is the odds ratio, which is the exponentiated logit coefficient. When the independent variable is continuous, for a one-unit increase from any value of  $x$  to the value of  $(x + 1)$ , the change in the odds is still the exponentiated logit coefficient. Since the 95% confidence intervals for the parameter  $\beta = \hat{\beta} \pm 1.96 * \text{SE}(\hat{\beta})$ , by exponentiating this form we obtain the corresponding confidence intervals for the odds ratio,  $\exp[\hat{\beta} \pm 1.96 * \text{SE}(\hat{\beta})]$ .

When the logit coefficient is positive, it indicates the relationship between the predictor variable and the logit function of the probability is positive. By exponentiating the logit coefficient, we get the odds ratio, which is larger than 1. This means that the odds of success or of having an event increases for a one-unit increase in the predictor variable.

When the logit coefficient is negative, it indicates that the relationship between the predictor variable and the logit function is negative. The exponentiated coefficient, the odds ratio, is less than 1. This means that the odds of success or of having an event decreases for a one-unit increase in the predictor variable.

When the logit coefficient equals 0, the odds ratio equals 1. This indicates that there is no relationship between the predictor and the odds of success.

### Interpreting an Odds Ratio as a Percentage Change in Odds

Another way of interpreting odds ratios is the percentage change in odds. It can be calculated by using  $(\text{odds ratio} - 1) \times 100\%$ . A positive percentage change in odds indicates there is an increase in the odds, whereas a negative percentage change corresponds to a decrease in the odds. A zero-percentage change indicates no change in odds at all. In other words, the predictor variable does not influence the odds of success.

For example, if an OR for a predictor variable equals 1.2, then the percentage change in the odds can be computed as follows:  $(1.2 - 1) \times 100\% = 20\%$ . This indicates that each one-unit increase in the predictor variable corresponds to an increase of 20% in the odds of success.

In another example, if OR = .80, then the percentage change in the odds is  $(0.80 - 1) \times 100\% = -20\%$ . Since the percentage change is negative, it indicates that for each one-unit increase in the predictor variable, there is a decrease of 20% in the odds of success.

## Interpreting Coefficients in Terms of Predicted Probabilities

In addition to odds ratios, we can interpret logistic regression models in terms of predicted probabilities or predictive margins. Recall that the logit is the natural logarithm of the odds, or log odds, and the odds of success is obtained by taking the exponential of the logit. By using the backward transformation method introduced previously, we can easily transform the odds back to the probability of success. The predicted probability can be calculated at specified values of a set of predictor variables. When a combination of the values of the predictor variable changes, the predicted probabilities are different.

## Interpreting Marginal Effects as Changes in Predicted Probabilities

A marginal effect is a change in a response variable related to the change in an independent variable. Mathematically, it is the partial derivative of the regression equation with respect to an individual predictor variable in the model. In a simple linear regression model, a marginal effect is the same as the regression coefficient. In logistic regression, a marginal effect is a change in the predicted probability with a change in the values of a predictor variable. When a predictor variable is binary with the values of 1 and 0, the marginal effect is a discrete change in the predicted probability; when a predictor variable is continuous, the marginal effect is an instantaneous rate of change in the predicted probability, which is a rate of change at a specific value of the predictor variable rather than the two values for a binary predictor variable. Therefore, a marginal effect for a continuous predictor variable in logistic regression is often an approximate change in the predicted probability with a one-unit change, or strictly speaking, it is a change in the predicted probability with a small change in that predictor variable. An average marginal effect (AME) is the mean value of the marginal effects at each value of a predictor variable. Marginal effects can be computed at means or at specified values of predictor variables, which are referred to as marginal effects at means (MEM) or marginal effects at representative values (MER), respectively.

We can also compute the predicted probabilities for a particular variable at given values. These predicted probabilities are also called predictive margins or adjusted probabilities. Changes in the predicted probabilities can be used to compute the marginal effects. For example, when a predictor variable is binary with two categories, the difference between the predicted probabilities is the marginal effect of that variable.

## 3.2 RESEARCH EXAMPLE AND DESCRIPTION OF THE DATA AND SAMPLE

---

*Research Problem and Questions:* In this example, we are interested in investigating whether the binary response variable, health status, can be predicted by four predictor variables, marital status, years of education, age, and gender. The research question is as follows: Which predictor variables are associated with the likelihood of having good health? In other words, can health status be predicted by the four preceding variables?



*Description of the Data and Sample:* The data for the following analyses were the General Social Survey 2016 (GSS 2016). The following are the variables:

- `healthy`: the recoded variable of health (health status) with 1 = good health and 0 = not good health
- `wrkfull`: working full-time with 1 = full-time and 0 = not full-time
- `maritals`: the recoded variable of marital (marital status) with 1 = currently married and 0 = not currently married
- `female`: recoded variable of sex with 1 = female and 0 = male
- `educ`: the highest education.

## 3.3 GENERALIZED LINEAR MODELS AND THE `glm()` FUNCTION

---

### 3.3.1 Generalized Linear Models and the `glm()` Function: An Introduction

Generalized linear models extend the linear regression model when the response variable is a noncontinuous variable, such as a binary, an ordinal, a nominal, or a count variable. When the response variable is continuous, the linear regression model is a special case of generalized linear models. In generalized linear models, the response variable follows a distribution from the exponential family, such as a normal distribution for a continuous response variable, a binomial distribution for a binary variable, and a Poisson distribution for a count variable. There are three common components in the framework of generalized linear models: the random component, the systematic component, and the link function.

First, a probability distribution of the response variable  $Y$  is the random component of the model. For example, the binomial distribution of a binary response variable is the random component in a logistic regression model.

Second, a generalized linear model also includes the systematic component, a linear combination of the predictor variables, which is referred to as the linear predictor with the form  $\alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$ . If we define the linear predictor as  $\eta$  (*eta*), the systematic component can be expressed as the form:  $\eta = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$ .

Third, a generalized linear model has a component called the link function to connect the linear predictor and the expected value  $E(Y)$  or the conditional mean  $\mu$  (mu) of the response variable. The link function is written as  $g(\mu)$ , which links the linear predictor  $\eta$  and the conditional mean  $\mu$  of the response variable. For example, the identity link,  $g(\mu) = \mu$ , is used for linear regression models, the logit link,  $g(\mu) = \ln(\mu/1-\mu)$ , is used for logistic regression models, and the natural log link,  $g(\mu) = \ln(\mu)$ , is used for Poisson regression and negative binomial regression models. These link functions are



the most commonly used canonical links which connect the linear predictor  $\eta$  with the natural parameter, the conditional mean  $\mu$ . Other link functions, such as the probit function and the complementary log-log, can also be used in models for binary, ordinal, and nominal response variables. Table 3.5 provides the types of outcome variables, the probability distributions, the link functions with names, and the corresponding models which are covered in this book.

**TABLE 3.5** Outcome Variables, Distributions, Link Functions, Function Names, and the Corresponding Models

Outcome Variable	Distribution	Link Function	Link Function Name	Model
Continuous	Gaussian	$\mu$	identity	Linear regression
Binary	Bernoulli or Binomial	$\ln(\mu/1 - \mu)$	logit	Logistic regression
Ordinal	Multinomial	$\ln\left(\frac{P(Y \leq j)}{P(Y > j)}\right)$	logit	Ordinal logistic regression (proportional odds model)
Ordinal	Multinomial	$\ln\left(\frac{P(Y = j)}{P(Y > j)}\right)$	logit	Ordinal logistic regression (continuation ratio odds model)
Ordinal	Multinomial	$\ln\left(\frac{P(Y = j + 1)}{P(Y = j)}\right)$	logit	Ordinal logistic regression (adjacent categories model)
Nominal	Multinomial	$\ln\left(\frac{P(Y = j)}{P(Y = J)}\right)$	logit	Multinomial logistic regression
Count	Poisson	$\ln(\mu)$	natural log	Poisson regression
Count	Negative binomial	$\ln\left(\frac{\alpha\mu}{1 + \alpha\mu}\right)$ (canonical form)	natural log	Negative binomial regression (canonical)
Count	Negative binomial	$\ln(\mu)$	natural log	Negative binomial (NB-2)

### 3.3.2 The `glm()` Function

The `glm()` function is normally used to fit generalized linear models where a binary logistic regression model belongs. The model formula syntax of `glm()` is similar to that of `lm()` introduced in the last chapter but includes the additional `family=` argument for the probability distribution of the outcome variable and the `link=` argument for the link function. The model formula in `glm()` specifies the dependent variable and the predictor variable(s), which are separated by the tilde (`~`). When there are multiple predictor variables in the formula, they are connected by plus (+) symbols.

In addition to the model formula, next, we also need to specify the probability distribution of the outcome variable with the `family=` argument. For example, the `family = binomial` argument is used for a binary outcome variable in logistic regression and `family = poisson` is specified for a count variable in Poisson regression. Other families include Gaussian, Gamma, inverse Gaussian, quasi, quasi-Poisson, and quasi-binomial. We also need to specify the link function associated with the `family` argument. For example, `(link = "logit")` stands for the logit link and `(link = "log")` is used for the log link. Several link functions can be used in the `binomial` family. In this chapter, we will only focus on the logit and probit links. Please also note that the current `glm()` function cannot be used to fit models for ordinal and nominal response variables, so other packages will be introduced in the following chapters.

Most extractor functions for the `lm()` function can also be used for the `glm()` function. We can use the `summary()` function to display the summary results of the fitted model, use the `coef()` function to extract the coefficients, use the `confint()` function to request the confidence intervals of the coefficients, and use the `anova()` function to conduct the likelihood ratio test for model comparison. In addition, we can the `exp(coef())` function to obtain the odds ratios. Other useful extractor functions include the `fitted()` function for creating the fitted values, the `residuals()` function for the residual values, the `predict()` function for the predicted values of an outcome variable, the `AIC()` function for the AIC statistic, and the `plot()` function for diagnostic plots.

## 3.4 SIMPLE LOGISTIC REGRESSION USING R

### 3.4.1 Simple Logistic Regression: R Syntax

In simple logistic regression, there is only one binary dependent variable with values of 1 and 0 and one independent variable, which can be either categorical or continuous. The command `glm(y ~ x, family = binomial(link = "logit"))` tells R to fit a simple logistic regression model predicting the binary dependent variable  $y$  with an independent variable  $x$  by specifying the binomial family and the logit link function. For more details on how to use this command, type `help(glm)` in the command prompt.

In the following example, the command `glm(healthy ~ wrkfull, data = chp3.lr, family = binomial(link = "logit"))` tells R to predict the dependent variable `healthy` from the independent variable `wrkfull`. The fitted model is named `LR.1`. The output is shown by the `summary(LR.1)` command.

```
> # Simple logistic regression
> LR.1 <- glm(healthy ~ wrkfull, data = chp3.lr, family = binomial(link = "logit"))
> summary(LR.1)
```

```
Call:
glm(formula = healthy ~ wrkfull, family = binomial(link = "logit"),
    data = chp2.lr)
```

```

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7721  -1.4222   0.6829   0.9509   0.9509

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.5592    0.0658   8.499  < 2e-16 ***
wrkfull        0.7778    0.1061   7.328  2.34e-13 ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2258.9  on 1872  degrees of freedom
Residual deviance: 2203.2  on 1871  degrees of freedom
AIC: 2207.2

Number of Fisher Scoring iterations: 4

```

### 3.4.2 Interpreting R Output

In the R output for simple logistic regression, the first part is the call, which shows the R syntax for the model. The second part shows the deviance residuals, which are similar to the residuals in a linear regression model. The minimum, first quarter, median, third quarter, and maximum values of the deviance residuals are shown here. The third part shows the coefficients table including the parameter estimates for the predictor variable and the intercept, their standard errors, the Wald  $z$  statistics, and the associated  $p$  values. The null hypothesis for the Wald test is that the coefficient of the predictor variable is 0, and the alternative hypothesis is that the coefficient of the predictor variable is significantly different from 0.

The Wald  $z$  statistic equals the parameter estimate divided by its standard error. For the predictor variable `wrkfull`,  $Wald\ z = .778/1.06 = 7.328$ . Some other statistical software packages report the Wald chi-square test statistic, which is the squared Wald  $z$  statistic. The associated  $p$  value,  $Pr(>|z|) < .001$ , so we rejected the null hypothesis. The rejection of the null hypothesis indicates that the predictor variable `wrkfull` is a significant predictor of the dependent variable `healthy`.

Finally, the fourth part of the output shows the fit statistics including the null deviance, the residual deviance, and the AIC. The null deviance is the deviance for the null model with the intercept only. The residual deviance is the deviance for the fitted model, which is defined as  $-2(\log \text{likelihood of the current model} - \log \text{likelihood of the$

saturated model). The difference between the residual deviance and the null deviance can be used to evaluate the significance of the fitted model. Please note that the residual deviance is not the same as the deviance residual. The deviance residual is the residual for each case, whereas the residual deviance is the total of the deviance residuals multiplied by 2.

### 3.4.3 Interpreting the Coefficients

The logit coefficients can also be obtained using `coef(LR.1)` and their confident intervals can be obtained with `confint(LR.1)`.

```
> coef(LR.1)
(Intercept)      wrkfull
  0.5592222      0.7778011

> confint(LR.1)
Waiting for profiling to be done...
              2.5 %      97.5 %
(Intercept)  0.4309678  0.6889906
wrkfull      0.5709823  0.9872410
```

Since the estimated intercept and coefficient are .559 and .778, respectively, the simple logistic regression model could be expressed as:

$$\text{logit}(p) = .559 + .778X$$

The regression coefficient of the `wrkfull` predictor is .778, which is also called the logit coefficient since we estimate the relationship between the predictor variable and the logit of the probability when the outcome variable `healthy` takes the value of 1.

The `confint(LR.2)` command produces the profile confidence intervals which are based on the chi-square distribution of the likelihood ratio test statistic. We can also use the `confint.default(LR.2)` command to obtain the standard confidence intervals, which are based on the Wald test statistic.

```
> confint.default(LR.1)
              2.5 %      97.5 %
(Intercept)  0.4302568  0.6881876
wrkfull      0.5697574  0.9858449
```

The results of the profile confidence intervals and the standard confidence intervals look similar. The profile confidence intervals are preferred when the sample size in the model is small. We will use the `confint()` function throughout the book.

### 3.4.4 Interpreting the Odds Ratio

When  $X = 0$ , the estimated logit( $p$ ) = .559. Since logit is log odds, exponentiating .559 gives us the odds of having good health. The odds of having good health for people who do not work full-time are  $\exp(.559) = 1.749$ .

When  $X = 1$ , the estimated logit( $p$ ) = .559 + .778 = 1.337. So the odds of having good health for people who work full-time are  $\exp(1.337) = 3.808$ .

Odds ratio =  $3.808/1.749 = 2.177$ , which is the ratio of odds of having good health for working full-time to the odds for not working full-time.

If we directly exponentiate the logit coefficient  $\exp(.778) = 2.177$ , then we get the same odds ratio as earlier. You get the following output displaying the odds ratio if you run the `exp(coef(LR.1))` command. The corresponding confidence intervals can be obtained using the `exp(confint(LR.1))` command.

```
> exp(coef(LR.1))
(Intercept)    wrkfull
  1.749311     2.176681

> exp(confint(LR.1))
Waiting for profiling to be done...

              2.5 %      97.5 %
(Intercept)  1.538746  1.991704
wrkfull      1.770005  2.683820
```

Odds ratio = 2.177. It is interpreted as follows: For each one-unit increase in the predictor variable, the odds of having good health increase by a factor of 2.177. Since the predictor variable `wrkfull` is also a binary variable, it can also be interpreted as the odds of having good health for the people who work full-time are 2.177 times the odds of having good health for those who do not work full-time.

### 3.4.5 Interpreting the Pseudo $R^2$

We use the `nagelkerke()` function in the `rcompanion` package (Mangiafico, 2021) to obtain the pseudo  $R^2$  statistics. The `rcompanion` package needs to be installed first by typing `install.packages("rcompanion")`. We then load it with `library(rcompanion)`. The output produced by the `nagelkerke(LR.1)` command is displayed as follows.

```
> # PseudoR2
> library(rcompanion)
> nagelkerke(LR.1)
$`Models`

Model: "glm, healthy ~ wrkfull, binomial(link = \"logit\"), chp3.1r"
```

```
Null: "glm, healthy ~ 1, binomial(link = \"logit\"), chp3.lr"
```

```
$Pseudo.R.squared.for.model.vs.null
```

	Pseudo.R.squared
McFadden	0.0246663
Cox and Snell (ML)	0.0293107
Nagelkerke (Cragg and Uhler)	0.0418352

```
$Likelihood.ratio.test
```

Df.diff	LogLik.diff	Chisq	p.value
-1	-27.86	55.72	8.3585e-14

```
$Number.of.observations
```

```
Model : 1873
Null   : 1873
```

```
$Messages
```

```
[1] "Note: For models fit with REML, these statistics are based on refitting with ML"
```

```
$Warnings
```

```
[1] "None"
```

McFadden's  $R^2 = 0.0247$ . It is the likelihood ratio  $R^2$ , which is normally written as  $R_L^2$ .

$$R_L^2 = 1 - \frac{-2LL_m}{-2LL_0} = 1 - (2,203.2/2,258.9) = 0.0247$$

```
> 1-2203.2/2258.9
[1] 0.02465802
```

The same results can be computed using the equations for the pseudo  $R^2$  introduced in the previous section. In the R syntax below, we first fit a null model with the intercept only and name the model object LR.0. We also create the following objects. LLM is the log-likelihood value for the single-predictor model and LL0 is the log-likelihood value for the null model. In addition, McFadden is the object name for the McFadden  $R^2$ , CS for the Cox and Snell  $R^2$ , and NG for the Nagelkerke  $R^2$ .

```

> LR.0 <- glm(healthy ~ 1, data = chp3.1r, family = binomial(link = "logit"))
> LLM <- logLik(LR.1)
> LL0 <- logLik(LR.0)
> McFadden <- 1 - (LLM/LL0)
> McFadden
'log Lik.' 0.02466634 (df=2)
> CS <- 1 - exp(2 * (LL0 - LLM) / 1873)
> CS
'log Lik.' 0.02931066 (df=1)
> NG <- CS / (1 - exp(2 * LL0 / 1873))
> NG
'log Lik.' 0.0418352 (df=1)

```

The results of the three pseudo  $R^2$  are the same as those produced by the `nagelkerke()` function.

### 3.4.6 AIC and BIC Statistics

AIC is reported by `summary(LR.1)`. The `AIC(LR.1)` command also reports the AIC statistic. The `BIC(LR.1)` command provides the BIC statistic.

```

> AIC(LR.1)
[1] 2207.209
> BIC(LR.1)
[1] 2218.28

```

In the output,  $AIC = 2,207.209$ . The same value can be easily derived if we use the following equation:

$$AIC = -2(LL_m - k) = (D_m + 2k) = 2,203.2 + 2 \times 2 = 2,207.2$$

BIC in the output is 2,218.28. It can be calculated using this equation:

$$BIC = D_m + \ln(n) \times k = 2,203.2 + \ln(1,873) \times 2 = 2,218.27$$

The AIC and BIC statistics themselves do not indicate whether the model fits the data well. They are useful for model comparison purposes, particularly when comparing non-nested models. More details will be provided in the multiple logistic regression section.

### 3.4.7 Testing the Overall Model Using the Likelihood Ratio Test

To test if the overall model is significant, we use the `anova(LR.1, update(LR.1, ~1), test = "Chisq")` command. The `anova()` function compares the deviance statistics of the fitted model `LR.1` and the null model using the chi-square test.

```

> LR.1 <- glm(healthy ~ wrkfull, data = chp3.lr, family = binomial(link = "logit"))
> anova(LR.1, update(LR.1, ~1), test = "Chisq")
Analysis of Deviance Table

Model 1: healthy ~ wrkfull
Model 2: healthy ~ 1

```

	Resid.Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	1871	2203.2			
2	1872	2258.9	-1	-55.72	8.358e-14 ***

```

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Deviance = 55.72, which is the likelihood ratio chi-square test statistic and can be written as  $LR \chi^2_{(1)} = 55.72$ . It is the difference in the  $-2 \log$  likelihood ( $-2LL$ ) between the current model, which contains the one predictor and the intercept, and the null or empty model, which contains only the intercept. The deviance of the current model was 2,203.2 and that of the null model was 2,258.9. Therefore, the difference in deviance was 55.72, with 1 degree of freedom, which is the difference in the number of predictors between these two models. Please note that the deviance is negative in the output since there is a decrease in the residual deviance from the null model to the one-predictor model.

The null hypothesis of the likelihood ratio chi-square test is that the logit coefficient of the predictor variable `wrkfull` is not significant or that the predictor variable does not contribute to the model.

The alternative hypothesis is that the coefficient of the predictor is significant in the model or that it significantly contributes to the model.

The associated  $p$  value with the likelihood ratio chi-square test  $\text{Prob}(>\text{chi}) < 0.0001$  indicates that the null hypothesis is rejected. Therefore, the overall model with one predictor is significant.

## 3.5 MULTIPLE LOGISTIC REGRESSION USING R

The multiple logistic regression model is simply an extension of the simple logistic regression model when there are two or more predictor variables in the model. The following equation is for the multiple logistic regression model:

$$\ln \frac{p(x)}{1 - p(x)} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p \quad (3.21)$$

where  $X_1, X_2, \dots, X_p$  are the predictor variables and  $\beta_1, \beta_2, \dots, \beta_p$  are the logit coefficients of these predictors.



The multiple logistic regression estimates the logit of the probability or log odds when the outcome variable equals 1 given a set of predictors, which can be either categorical or continuous. The odds of success or of having an event in multiple regression is the ratio of probability of success to the probability of failure given a set of predictor variables.

### 3.5.1 Interpretation of Model Parameter Estimates and Odds Ratios in Multiple Logistic Regression

The estimated regression coefficients are the logit coefficients, which are the coefficients on the scale of logit or log odds. The logit coefficient can be interpreted as the change in the predicted logit or the log odds for a one-unit increase in the predictor variable when holding other predictors constant or controlling for the effects of other predictors.

The exponentiated logit coefficients are the odds ratios. The odds ratio of each predictor can be interpreted as the change in the odds for a one-unit change in the dependent variable when holding other predictors constant.

When an odds ratio is larger than 1, the odds of success or of having an event increase for a one-unit increase in the predictor variable when controlling for other predictors.

When an odds ratio is less than 1, the odds of success or of having an event decrease for a one-unit increase in the predictor variable when controlling for other predictors.

When an odds ratio equals 1, there is no relationship between the predictor and the odds of success when holding other variables constant.

### 3.5.2 Model Fitting Based on the Likelihood Ratio Test and Information Criteria Statistics

Just like linear regression models, models in logistic regression can be fitted from a simple logistic regression model to a more complex model to the full model. A baseline model can be the starting point for model building. A series of nested models can be compared by using the likelihood ratio test or the deviance difference test. Models are nested when one model is a special case of the other. As explained in the Goodness-of-Fit Statistics section, the difference in deviance between nested models has an approximately chi-square distribution with the degrees of freedom equal to the difference in the number of parameters between these two models. It is often expressed as  $G = \text{deviance for the reduced model} - \text{deviance for the full model}$  or as  $D_{\text{Reduced}} - D_{\text{Full}}$ . If the likelihood ratio chi-square test or the deviance test is significant, then we reject the null hypothesis and conclude that one model fits the data better than the other model.

The result of the likelihood ratio test is valid when the competing models are fitted on the same data. In real data analysis, it is common to see that the sample size varies among the nested models. If there is a great difference between the sample sizes of fitted models, for example, if some variables have more missing values, then the results are questionable. Under this circumstance, two information criteria measures, the AIC and BIC statistics, are more appropriate for model comparisons. They are commonly used

to compare non-nested models. The smaller the AIC and BIC statistics, the better fit the model.

### 3.5.3 The `glm()` Function for Multiple Logistic Regression

We still use the `glm()` function for multiple logistic regression. In the model formula, the dependent variable `healthy` and the four predictor variables are separated by the tilde (`~`). The four predictor variables include `wrkfull`, `maritals`, `female`, and `educ`, which are connected by plus (`+`) symbols. We also specify the data and the family arguments.

In the following example, the `glm(healthy ~ wrkfull + maritals + female + educ, data = chp3.lr, family = binomial(link = "logit"))` command tells R to predict the dependent variable `healthy` from the four independent variables. The fitted model is named `LR.2`. The following output is shown by the `summary(LR.2)` command.

```
> # Multiple logistic regression
> LR.2 <- glm(healthy ~ wrkfull + maritals + female + educ, data = chp3.lr, family =
binomial(link = "logit"))
> summary(LR.2)

Call:
glm(formula = healthy ~ wrkfull + maritals + female + educ, family = binomial(link =
"logit"),
    data = chp3.lr)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3201  -1.1877   0.6439   0.8652   1.7620

Coefficients:
            Estimate      Std. Error  z value Pr(>|z|)
(Intercept)  -1.82016     0.26294  -6.922  4.45e-12 ***
wrkfull       0.65505     0.11112   5.895  3.74e-09 ***
maritals     0.36738     0.10943   3.357  0.000787 ***
female       0.18217     0.10807   1.686  0.091868 .
educ         0.16184     0.01881   8.605  < 2e-16 ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2258.9  on 1872  degrees of freedom
Residual deviance: 2103.4  on 1868  degrees of freedom
AIC: 2113.4

Number of Fisher Scoring iterations: 4
```

### 3.5.4 Interpreting R Output

In the R output for multiple logistic regression, the first part is the call, which shows the R syntax for the model. The second part shows the minimum, first quarter, median, third quarter, and maximum value of the deviance residuals. The third part shows the coefficients table including the parameter estimates for the four predictor variables and the intercept, their standard errors, the Wald  $z$  statistics, and the associated  $p$  values.

For the predictor variable `wrkfull`, Wald  $z = 5.895$ . The associated  $p$  value,  $\Pr(>|z|) < .001$ , so we rejected the null hypothesis. The rejection of the null hypothesis indicates that the predictor variable `wrkfull` is a significant predictor of the dependent variable `healthy`.

For the predictor variable `maritals`, Wald  $z = 3.357$ . The associated  $p$  value,  $\Pr(>|z|) < .001$ , so we rejected the null hypothesis. For the predictor variable `educ`, the Wald  $z = 8.605$ . The associated  $p$  value,  $\Pr(>|z|) < .001$ , so we also reject the null hypothesis. Therefore, `maritals` and `educ` are significant predictors of the dependent variable.

For the other new predictor variable `female`, the Wald  $z = 1.686$ . The associated  $p$  value  $\Pr(>|z|) = .092$ , so we fail to reject the null hypothesis and conclude that there is no significant effect of `female` on the outcome variable. In other words, whether a person is a female or male does not significantly predict whether that person has good health.

Finally, the fourth part of the output shows the fit statistics including the null deviance, the residual deviance, and the AIC.

### 3.5.5 Interpreting the Coefficients

The logit coefficients can also be obtained using the `coef(LR.2)` command and their confident intervals can be obtained with the `confint(LR.2)` command.

```
> coef(LR.2)
(Intercept)      wrkfull      maritals      female      educ
-1.8201572      0.6550481      0.3673780      0.1821725      0.1618404

> confint(LR.2)
Waiting for profiling to be done...
              2.5 %      97.5 %
(Intercept) -2.34090516  -1.3094672
wrkfull      0.43826244   0.8740228
maritals     0.15362395   0.5827676
female      -0.02966614    0.3941360
educ        0.12540105    0.1991806
```

The regression coefficients of the predictors in multiple logistic regression are the logit coefficients since they are on the scale of the logit or log odds. They are the partial

effects of the predictor variables on the logit or log odds of the outcome variable (having good health). Each logit coefficient can be interpreted as the change in the logit or log odds of having good health for each one-unit increase in the predictor when controlling for the effects of other variables. A positive logit coefficient indicates there is an increase in the logit coefficient for a one-unit increase in the predictor, whereas a negative logit coefficient indicates that there is a decrease in the logit coefficient for a one-unit increase in the predictor when controlling for other predictors.

### 3.5.6 Interpreting the Odd Ratios

You get the following output displaying the odds ratios if you run the `exp(coef(LR.2))` command. The corresponding confidence intervals can be obtained using the `exp(confint(LR.2))` command.

```
> exp(coef(LR.2))
(Intercept)      wrkfull      marital      female      educ
0.1620003      1.9252351      1.4439437      1.1998212      1.1756726

> exp(confint(LR.2))
Waiting for profiling to be done...
                2.5 %      97.5 %
(Intercept)      0.09624049      0.2699638
wrkfull          1.55001164      2.3965322
maritals         1.16605232      1.7909884
female           0.97076958      1.4831023
educ             1.13360300      1.2204023

> cbind(exp(coef(LR.2)), exp(confint(LR.2)))
Waiting for profiling to be done...
                2.5 %      97.5 %
(Intercept)      0.1620003      0.09624049      0.2699638
wrkfull          1.9252351      1.55001164      2.3965322
maritals         1.4439437      1.16605232      1.7909884
female           1.1998212      0.97076958      1.4831023
educ             1.1756726      1.13360300      1.2204023
```

For `wrkfull`, odds ratio = 1.925. The odds of having good health for the people who work full-time are 1.925 times the odds of having good health for those who do not work full-time.

For `maritals`, odds ratio = 1.444, which is larger than 1. This indicates that the odds of having good health for the married were 1.444 times the odds for the unmarried.

For `educ`, odds ratio = 1.176, for each one-unit increase in education, the odds of having good health increase by a factor of 1.176. Another interpretation is the percentage change in odds. It can be calculated by using  $(\text{odds ratio} - 1) \times 100\%$ . In this example,  $(1.176 - 1) \times 100\% = 17.6\%$ . For each one-unit increase in education, there is an increase of 17.6% in the odds of having good health.

For female, odds ratio = 1.200,  $p > .05$ , which is not significant. When an odds ratio takes a value of 1, it means that there is no significant effect of the predictor variable on the odds of success when holding other variables constant. Therefore, being a female or not does not significantly influence whether that person has good health.

### 3.5.7 Interpreting the Pseudo $R^2$

The pseudo  $R^2$  statistics for the multiple logistic regression model are obtained using the `nagelkerke (LR.2)` command. The output is displayed as follows.

```
> # PseudoR2 (LR.2)
> nagelkerke (LR.2)
$`Models`

Model: "glm, healthy ~ wrkfull + marital + female + educ, binomial(link = \"logit\"), chp3.1r"
Null: "glm, healthy ~ 1, binomial(link = \"logit\"), chp3.1r"

$Pseudo.R.squared.for.model.vs.null

                                Pseudo.R.squared
McFadden                        0.0688652
Cox and Snell (ML)              0.0796992
Nagelkerke (Cragg and Uhler)    0.1137550

$Likelihood.ratio.test
Df.diff   LogLik.diff   Chisq   p.value
  -4      -77.781     155.56  1.3082e-32

$Number.of.observations

Model   :   1873
Null    :   1873

$Messages
[1] "Note: For models fit with REML, these statistics are based on refitting with ML"

$Warnings
[1] "None"
```

The McFadden  $R^2 = 0.069$ , the Cox Snell  $R^2 = 0.080$ , and the Nagelkerke  $R^2 = 0.114$ .

These three pseudo  $R^2$  statistics are larger than those in the previous simple logistic regression model. (McFadden's  $R^2 = .025$  for the one-predictor model.) It seems that the full model fits better than the simple logistic regression model, but we also need to look at other fit statistics.

Please note that a pseudo  $R^2$  measure should not be interpreted as variance explained by the predictors in the model as its analog in multiple linear regression. When it is used for model comparisons, instead of depending on this single statistic, it should be used together with other fit statistics.

### 3.5.8 AIC and BIC Statistics

The AIC and BIC statistics are obtained using `AIC(LR.2)` and `BIC(LR.2)`, respectively.

```
> AIC(LR.2)
[1] 2113.367
> BIC(LR.2)
[1] 2141.044
```

To compare the AIC and BIC statistics of the two fitted model, we use `AIC(LR.1, LR.2)` and `BIC(LR.1, LR.2)`.

```
> AIC(LR.1, LR.2)
      df      AIC
LR.1   2  2207.209
LR.2   5  2113.367

> BIC(LR.1, LR.2)
      df      BIC
LR.1   2  2218.280
LR.2   5  2141.044
```

The AIC and BIC statistics decrease by 93.842 and 77.236, respectively. Since smaller AIC and BIC statistics means a better model fit, the results support that the multiple logistic regression model fits the data better.

### 3.5.9 Hosmer–Lemeshow Goodness-of-Fit Statistic

We use the `hoslem.test()` function in the `ResourceSelection` package (Lele et al., 2019) to obtain the Hosmer–Lemeshow goodness-of-fit statistic. The `ResourceSelection` package needs to be installed first by typing `install.packages("ResourceSelection")`. We then load it with `library(ResourceSelection)`. The output produced by the `hoslem.test(healthy, fitted(LR.2), g = 10)` syntax is displayed as follows. The syntax includes the outcome variable, `healthy`, the estimated probabilities, `fitted(LR.2)`, and the number of groups, `g = 10`.

```

> # Hosmer-Lemeshow goodness-of-fit test
> library(ResourceSelection)
> hoslem.test(healthy, fitted(LR.2), g = 10)

Hosmer and Lemeshow goodness of fit (GOF) test

data: healthy, fitted(LR.2)
X-squared = 7.9332, df = 8, p-value = 0.44

> hlt2<-hoslem.test(healthy, fitted(LR.2), g = 10)
> cbind(hlt2$observed,hlt2$expected)

```

	y0	y1	yhat0	yhat1
[0.212,0.53]	109	82	102.48990	88.5101
(0.53,0.575]	79	111	82.67738	107.3226
(0.575,0.652]	79	131	79.11126	130.8887
(0.652,0.685]	60	132	62.82160	129.1784
(0.685,0.723]	47	135	52.67628	129.3237
(0.723,0.758]	54	131	46.16049	138.8395
(0.758,0.79]	40	144	39.57522	144.4248
(0.79,0.833]	29	176	37.20245	167.7975
(0.833,0.857]	23	127	22.74877	127.2512
(0.857,0.932]	25	159	19.53663	164.4634

In the output, the Hosmer–Lemeshow chi-square test has a value of 7.933, with the degrees of freedom equal to 8. The associated  $p$  value is .44, which is not significant. Therefore, the model fits the data well.

To see the estimated probabilities by the 10 groups, we use the `cbind(hlt2$observed, hlt2$expected)` command. The first column lists the estimated probabilities for these groups, which are ordered in sequence. The second column `y0` lists the number of observations for the outcome variable with the value of 0, and the third column `y1` is the number of observations for the outcome variable with the value of 1. The last two columns are `yhat0` and `yhat1`, respectively, which are the expected number of frequencies for no event and the event occurring. Within each group, the frequencies of the observed cases are compared with the expected frequencies, and we would like to see small discrepancies between them. This test follows a chi-square distribution with the degrees of freedom equal to the number of groups  $-2$ . A nonsignificant  $p$  value indicates that the model fits the data well since there is no significant difference between the observed and expected data.

### 3.5.10 Testing the Overall Model Using the Likelihood Ratio Test

To test if the overall model is significant, we use the `anova(LR.2, update(LR.2, ~1), test = "Chisq")` command. The `anova()` function compares the deviance statistics of the fitted model LR.2 and the null model using the chi-square test.

```

> LR.2 <- glm(healthy ~ wrkfull + maritals + female + educ, data = chp3.lr, family = binomial(link =
"logit"))
> anova(LR.2, update(LR.2, ~1), test = "Chisq")
Analysis of Deviance Table

Model 1: healthy ~ wrkfull + maritals + female + educ
Model 2: healthy ~ 1

```

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	1868	2103.4			
2	1872	2258.9	-4	-155.56	< 2.2e-16 ***

```

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

The deviance of the current model was 2,103.4 and that of the null model was 2,258.9. Therefore, the difference in deviance was 155.56 with 4 degrees of freedom. The deviance difference is the likelihood ratio chi-square test statistic. It is the difference in the  $-2$  log likelihood ( $-2LL$ ) between the current model, which contains the four predictors and the intercept, and the null or empty model, which contains only the intercept.

The associated  $p$  value with the likelihood ratio chi-square test  $\text{Prob}(>\text{chi}) < 0.0001$  indicates that the null hypothesis is rejected. Therefore, the overall model with the four predictors is significant.

### 3.5.11 Model Comparison Using the Likelihood Ratio Test

To identify which model fits the data better, the likelihood ratio test or the deviance difference test can be used. Recall that this test compares the reduced model, which contains less parameters, and the full model, which contains all parameters. The difference in deviance is often expressed as  $G = \text{Deviance for the reduced model} - \text{Deviance for the full model}$  or as  $D_{\text{Reduced}} - D_{\text{Full}}$ . The difference in deviance between nested models has a chi-square distribution with the degrees of freedom equal to the difference in the number of parameters between these two models.

The `anova()` function is used for the likelihood ratio test or the deviance difference test. Next, we compare the simple logistic regression model and the multiple logistic regression model with the `anova(LR.1, LR.2, test = "Chisq")` command.

```

> anova(LR.1, LR.2, test = "Chisq")
Analysis of Deviance Table

Model 1: healthy ~ wrkfull
Model 2: healthy ~ wrkfull + maritals + female + educ

```

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	1871	2203.2			
2	1868	2103.4	3	99.842	< 2.2e-16 ***

```

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```



The likelihood ratio chi-square test  $\chi^2_{(3)} = 99.842$ ,  $p < .001$ . The same result can be obtained if you compute it using the following equation:

$$G = D_{\text{Reduced}} - D_{\text{Full}} = 2203.2 - 2103.4 = 99.80, \text{ df} = 1871 - 1868 = 3$$

### 3.5.12 Interpreting the Marginal Effects in Logistic Regression

We load the `margins` package (Leeper, 2021) with the `library(margins)` command and compute the average marginal effects with the `margins(LR.2)` command. This package needs to be installed first with the `install.packages("margins")` command. The results are named `LR.2`, and the summary of the results is obtained with `summary(marg2)` as follows.

```
> library(margins)
> marg2<-margins(LR.2)
> summary(marg2)
```

factor	AME	SE	z	p	lower	upper
educ	0.0307	0.0033	9.2173	0.0000	0.0241	0.0372
female	0.0345	0.0204	1.6896	0.0911	-0.0055	0.0745
maritals	0.0696	0.0205	3.3897	0.0007	0.0293	0.1098
wrkfull	0.1241	0.0204	6.0787	0.0000	0.0841	0.1641

The average marginal effect for `educ` is .031. Since `educ` is continuous, its marginal effect is an instantaneous rate of change in the probability of having good health with a small change in education. The result indicates that on average the predicted probability of having good health increases by 3.1% for a one-unit increase in education when holding all other predictors constant.

The average marginal effect for `maritals` is .070. Since `maritals` is binary, the marginal effect is a discrete change when its value moves from 0 to 1. This indicates that on average the predicted probability of having good health for the married is 7% higher than that for the unmarried when holding the other predictors constant. The other two predictor variables are also binary, so the interpretation is similar.

### 3.5.13 Computing the Predicted Probabilities With the `predict()` Function

We can use the `predict()` function to compute the predicted probabilities for a particular variable at given values. For example, we would like to compute the predicted probabilities for `educ` at the specified values of 12, 14, and 16 when holding the other predictor variables at their means. We first create a data frame with the `data.frame()` function and then apply the `predict()` function. In the `data.frame()` function, `educ = c(12, 14, 16)` specifies the values of `educ`; `maritals = rep(mean(maritals), 3)` repeats the mean of `maritals` three times; `wrkfull = rep(mean(wrkfull), 3)` repeats the mean of

wrkfull three times; and female = rep(mean(female), 3) repeats the mean of female three times. The created data frame is assigned to an object named newdf.

```
> newdf <- data.frame(educ = c(12,14,16),
+                     maritals=rep(mean(maritals), 3),
+                     wrkfull=rep(mean(wrkfull), 3),
+                     female=rep(mean(female), 3))
> newdf
```

	educ	maritals	wrkfull	female
1	12	0.4372664	0.467165	0.5563267
2	14	0.4372664	0.467165	0.5563267
3	16	0.4372664	0.467165	0.5563267

In the predict() function, we first specify the model object LR.2 and then the newdata = newdf argument, followed by the type = "response" argument for the predicted probabilities. The predicted probabilities labeled pred.prob are provided in the data frame named newdf.

```
> newdf[,c('pred.prob')] <- predict(LR.2, newdata=newdf, type="response")
> newdf
```

	educ	maritals	wrkfull	female	pred.prob
1	12	0.4372664	0.467165	0.5563267	0.6659520
2	14	0.4372664	0.467165	0.5563267	0.7337268
3	16	0.4372664	0.467165	0.5563267	0.7920447

The same results can be obtained by using the ggpredict() function in the ggeffects package in the following section.

### 3.5.14 Computing the Predicted Probabilities With the ggpredict() Function in the ggeffects Package

The ggpredict() function in the ggeffects package (Lüdtke, 2018b) is used to compute the predicted probabilities when the outcome variable  $Y = 1$  for predictor variables at specified values. The ggeffects package needs to be installed first by typing install.packages("ggeffects"). We then load it with library(ggeffects). In the first example, we compute the predicted probabilities of being healthy (i.e.,  $Y = 1$ ) for educ at the values of 12, 14, and 16 when holding the other three predictor variables constant at their means. Please note that the predictor variables need to be numeric so that they can be held constant at their means. The syntax is as follows: margins <- ggpredict(LR.2, terms = "educ[12, 14, 16]"). In the ggpredict() function, LR.2 is the fitted model and the terms = "educ[12, 14, 16]" option specifies the predictor variable educ at the values of 12, 14, and 16. When there are more than one variable, the terms option can specify

up to four variables, including the second to fourth grouping variables. In this example, we include only one variable in the `terms` option. The output is assigned to an object named `margins`. To request the standard errors of the predicted probabilities, we can use either the `as.data.frame()` or the `sqrt(diag(vcov()))` function. The `vcov()` function computes the variance–covariance matrix, so the standard errors can be obtained by taking the square root of the variances, which are on the diagonal of the matrix.

```
> library(ggeffects)
> margins <- ggpredict(LR.2, terms = "educ[12, 14, 16]")
> margins

# Predicted values of healthy

educ | Predicted | 95% CI
-----|-----|-----
12 | 0.67 | [0.64, 0.69]
14 | 0.73 | [0.71, 0.75]
16 | 0.79 | [0.77, 0.82]

Adjusted for:
* wrkfull = 0.47
* marital = 0.44
* female = 0.56
> as.data.frame(margins)
  x predicted std.error conf.low conf.high group
1 12 0.6659520 0.05899403 0.6397564 0.6911624 1
2 14 0.7337268 0.05581622 0.7118151 0.7545469 1
3 16 0.7920447 0.07470175 0.7668988 0.8151319 1

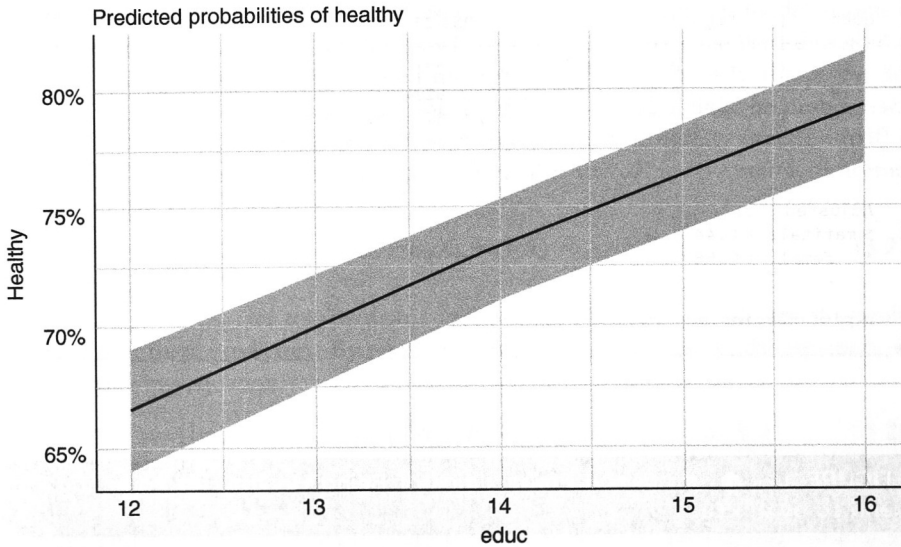
> sqrt(diag(vcov(margins)))
      1      2      3
0.05899403 0.05581622 0.07470175

> plot(margins)
```

When `educ` equals 12, 14, and 16, and the other three predictor variables are held at their means, the predicted probabilities of being healthy (i.e.,  $Y = 1$ ) are .666, .734, and .792, respectively. The results are the same as those obtained in the last section. They are plotted using the `plot(margins)` command. Figure 3.1 shows the predicted probabilities of being healthy (i.e.,  $Y = 1$ ) for `educ` at 12, 14, and 16. With an increase in years of education, the probabilities of having good health ( $Y = 1$ ) increase.

We can also compute the predicted probabilities for a continuous variable at given values by different groups. In the third example, we compute the predicted probabilities of being healthy (i.e.,  $Y = 1$ ) for `educ` at the values of 12, 14, and 16 by the two groups in `wrkfull` when holding other variables at their means. The syntax is as

**FIGURE 3.1** Estimated Probabilities When  $Y = 1$  for educ at 12, 14, and 16 With Others Fixed at Their Means



follows: `margins.ew <- ggpredict(LR.2, terms = c("educ[12, 14, 16]", "wrkfull"))`. In the `ggpredict()` function, the `terms = c("educ [12, 14, 16]", "wrkfull")` option specifies both `educ` and `wrkfull`, with the latter as the grouping variable. The output is assigned to an object named `margins.ew` and is plotted with the `plot(margins.ew)` command.

```
> margins.ew <- ggpredict(LR.2, terms = c("educ[12, 14, 16]", "wrkfull"))
> margins.ew
```

```
# Predicted values of healthy
```

```
# wrkfull = 0
```

educ	Predicted	95% CI
12	0.59	[0.56, 0.63]
14	0.67	[0.64, 0.70]
16	0.74	[0.70, 0.77]

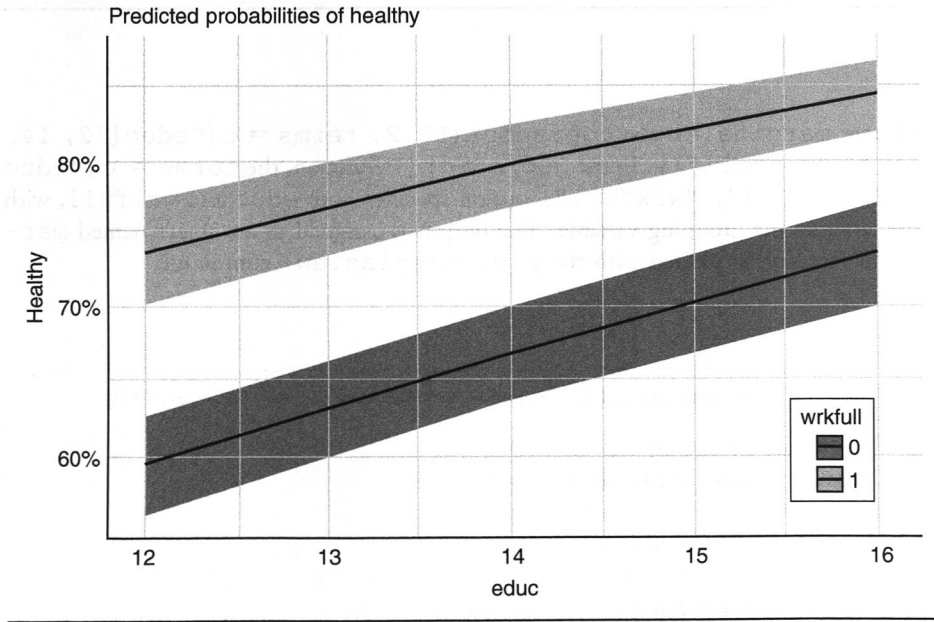
```
# wrkfull = 1

educ | Predicted | 95% CI
-----|-----|-----
12 | 0.74 | [0.70, 0.77]
14 | 0.80 | [0.77, 0.82]
16 | 0.84 | [0.82, 0.87]

Adjusted for:
* marital = 0.44
* female = 0.56

> plot(margins.ew)
```

**FIGURE 3.2** Predicted Probabilities of Being Healthy for educ at 12, 14, and 16 by wrkfull



As shown in the graph, the predicted probabilities of having good health ( $Y = 1$ ) increase with an increase in years of education and the probability for the people who work full-time is higher than the probability for those not working full-time.

## 3.6 PROBIT REGRESSION USING R

In addition to the logistic regression models introduced above, probit regression is also a popular method for analyzing binary response variables. The probit model assumes a standard cumulative normal distribution with a mean of 0 and a standard deviation of 1. While the logit link function is used for logistic regression models, the probit link function is used for the probit regression models. The probit link, or probit transformation, can be expressed as the inverse of the cumulative density function (cdf) of the standard normal distribution  $\Phi^{-1}(\pi)$ , where  $\Phi$  is the cdf for the standard normal. The model can be expressed as:

$$\Phi^{-1}(\pi) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p \quad (3.22)$$

where  $X_1, X_2, \dots, X_p$  are the predictor variables and  $\beta_1, \beta_2, \dots, \beta_p$  are the probit coefficients of these predictors. By taking the inverse on both sides of the equation, we get the following equation:

$$\pi(Y = 1 | x_1, x_2 \dots x_p) = \Phi(\alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p) \quad (3.23)$$

where  $\Phi(\cdot)$  denotes the cdf for the standard normal and a linear combination of the intercept and a set of predictors that are specified within the parentheses. The probit coefficients are  $z$  scores or standard normal scores.

### 3.6.1 Interpretation of Model Parameter Estimates in Probit Regression

First, the estimated regression coefficients are the probit coefficients, which can be interpreted as the change in the predicted probit for a one-unit increase in the predictor variable when holding the other predictors constant. Second, since probit regression is based on the cumulative normal distribution  $\Phi$ , the regression coefficient can also be interpreted as the change in the  $z$  scores for each one-unit increase in the predictor variable. Recall that the change in  $z$  scores is the number of standard deviations from the mean of a variable. Third, as with logistic regression, we can estimate marginal effects in probit regression. A marginal effect can be interpreted as the change in the predicted probability for a one-unit change in a predictor variable. Fourth, we can interpret probit coefficients in terms of predicted probabilities at representative values of a set of predictor variables.

### 3.6.2 The `glm()` Function for Multiple Probit Regression

In the following example, the `glm(healthy ~ wrkfull + marital + female + educ, data = chp3.lsr, family = binomial(link = "probit"))` command tells R to predict the dependent variable `healthy` from the four independent variables with the probit link function. The fitted model is named `PR.2`. The following output is shown by the `summary(PR.2)` command.

```

> # Probit regression
> PR.2 <- glm(healthy ~ wrkfull + maritals + female + educ, data = chp3.1r, family =
binomial(link = "probit"))
> summary(PR.2)

Call:
glm(formula = healthy ~ wrkfull + maritals + female + educ, family = binomial(link =
"probit"),
    data = chp2.1r)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3716  -1.2015   0.6466   0.8735   1.7227

Coefficients:
              Estimate      Std. Error    z value    Pr(>|z|)
(Intercept)  -1.04382      0.15487    -6.740    1.58e-11 ***
wrkfull       0.38772      0.06535     5.933    2.98e-09 ***
maritals      0.21948      0.06463     3.396    0.000684 ***
female        0.10578      0.06423     1.647    0.099553 .
educ          0.09425      0.01096     8.599    < 2e-16 ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2258.9  on 1872  degrees of freedom
Residual deviance: 2105.0  on 1868  degrees of freedom
AIC: 2115

Number of Fisher Scoring iterations: 4

```

The probit coefficients can also be obtained using the `coef(PR.2)` command and their confident intervals can be obtained with the `confint(PR.2)` command. The output is omitted here.

### 3.6.3 Interpreting Probit Coefficients in R Output

The coefficients table displays the probit coefficients for the four predictor variables and the intercept, their standard errors, the Wald  $z$  statistics, and the associated  $p$  values.

For the predictor variable `wrkfull`, the probit coefficient  $\beta = .388$ , which is positive. This indicates that working full-time increases the probability of having good health when holding all the other predictors constant. It can also be interpreted that working full-time increases the  $z$  score by .388. The probit coefficients of the other three predictors can be interpreted in a similar way.





We can also create the table in the HTML format and copy it into Microsoft Word. The syntax is as follows: `stargazer(LR.1, LR.2, type = "html", align = TRUE, out = "lr2mod.htm")`. It produces Table 3.6, as shown here in its original format, presenting the results of both the single-predictor and multiple-predictor logistic regression models.

**TABLE 3.6** Results of the Logistic Regression Models: Single-Predictor and Multiple-Predictor Models (Shown in Original Format Generated by R)

	<i>Dependent variable:</i>	
	Healthy	
	(1)	(2)
Wrkfull	0.778*** (0.106)	0.655*** (0.111)
Maritals		0.367*** (0.109)
Female		0.182* (0.108)
Educ		0.162*** (0.019)
Constant	0.559*** (0.066)	-1.820*** (0.263)
Observations	1,873	1,873
Log Likelihood	-1,101.605	-1,051.684
Akaike Inf. Crit.	2,207.209	2,113.367

Note:  
 \* $p < 0.1$   
 \*\* $p < 0.05$   
 \*\*\* $p < 0.01$

### 3.8 REPORTING THE RESULTS

When you summarize the results of logistic regression models, it is good practice to describe the statistical method you use for data analysis, the dependent variable and the independent variables in the models, and your research hypothesis or the purpose of your study. Also, report the likelihood ratio test statistic, the degrees of freedom, and the associated  $p$  value. Based on the likelihood ratio test result, you may discuss whether

the fitted model fits better than the null model containing only the intercept. If more than one model is fitted, then the deviance statistics for each model may be presented, and the results of the likelihood ratio test comparing deviance statistics between nested models may be provided.

In the body of the text, you may interpret the parameter estimates or the odds ratios for the predictor variables.

You could have a table containing all the parameter estimates, their standard errors,  $p$  values, and odds ratios. A summary of major fit statistics also might be included in the table, which could include the likelihood ratio test statistic with its degrees of freedom; a couple of pseudo  $R^2$  values, such as the likelihood ratio  $R^2$  and Cox and Snell's  $R^2$ ; the deviance statistic; and the AIC and BIC statistics. If more than one model is fitted, then the results of all the competing models from the simple model to the full model can be presented in the table. For detailed guidelines and recommendations on reporting the results of logistic regression, see Peng et al. (2002). In addition, O'Connell and Amico (2010) provided a list of key elements of logistic regression that should be addressed when judging a manuscript from a reviewer's perspective. Huck (2012) also provided examples on reporting the results of logistic regression from published articles. The following is an example of summarizing results from the multiple logistic regression model.

The multiple logistic regression analysis was conducted to estimate the probability of having good health from four predictor variables. The dependent variable was having good health or not, and the independent variables were working full-time, marital status, years of education, and gender. The likelihood ratio chi-square test statistic for the four-predictor model  $LR \chi^2_{(4)} = 155.56$ ,  $p < .001$ , indicated that the overall model with all four predictors was significant. When comparing the multiple logistic regression model with the simple logistic regression model, the likelihood ratio chi-square test  $\chi^2_{(3)} = 99.842$ ,  $p < .001$ , indicated the full model had a better fit. Table 3.6 presents the logit coefficients and standard errors of the full model. The results are interpreted in terms of odds ratios.

For `wrkfull`, odds ratio = 1.925. The odds of having good health for the people who work full-time are 1.925 times the odds of having good health for those who do not work full-time.

For `maritals`, odds ratio = 1.444, which is larger than 1. This indicates that the odds of having good health for the married were 1.444 times the odds for the unmarried.

For `educ`, odds ratio = 1.176, for each one-unit increase in education, the odds of having good health increase by a factor of 1.176.

For `female`, odds ratio = 1.200,  $p > .05$ , which is not significant. This indicates that being a female or not does not significantly influence whether that person has good health.

## 3.9 SUMMARY OF R COMMANDS IN THIS CHAPTER

---

```

# Chap 3 R Script
# Remove all objects
rm(list = ls(all = TRUE))

# The following user-written packages need to be installed first by using
install.packages(" ") and then by loading it with library()
# library(foreign)
# library(rcompanion)
# library(ResourceSelection)
# library(margins)
# library(ggeffects) # It is already installed for Chapter 2
# library(stargazer) # It is already installed for Chapter 2

# Import GSS 2016 Stata data file
library(foreign)
chp3.lr <- read.dta("C:/CDA/gss2016.dta")
chp3.lr$educ <- as.numeric(chp3.lr$educ)
chp3.lr$wrkfull <- as.numeric(chp3.lr$wrkfull)
chp3.lr$maritals <- as.numeric(chp3.lr$maritals)
attach(chp3.lr)
head(chp3.lr)
table(healthy)
table(wrkfull)

# Cross tabulation
tab <- table(healthy, wrkfull)
summary(tab)
tab
ftable(tab)
addmargins(tab)

# Simple logistic regression
LR.1 <- glm(healthy ~ wrkfull, data = chp3.lr, family = binomial(link = "logit"))
summary(LR.1)
coef(LR.1)
confint(LR.1)
confint.default(LR.1)
exp(coef(LR.1))
exp(confint(LR.1))

library(rcompanion)
nagelkerke(LR.1)
1-2203.2/2258.9
LR.0 <- glm(healthy ~ 1, data = chp3.lr, family = binomial(link = "logit"))
LLM <- logLik(LR.1)
LL0 <- logLik(LR.0)
McFadden <- 1-(LLM/LL0)

```

```

McFadden
CS <- 1-exp(2*(LL0-LLM)/1873)
CS
NG <- CS/(1-exp(2*LL0/1873))
NG

AIC(LR.1)
BIC(LR.1)

anova(LR.1, update(LR.1, ~1), test="Chisq")

# Multiple logistic regression
LR.2 <- glm(healthy ~ wrkfull + maritals + female + educ, data = chp3.lr, family =
binomial(link = "logit"))
summary(LR.2)
coef(LR.2)
confint(LR.2)
exp(coef(LR.2))
exp(confint(LR.2))
cbind(exp(coef(LR.2)), exp(confint(LR.2)))

nagelkerke(LR.2)
1-2103.4/2258.9
AIC(LR.2)
BIC(LR.2)
AIC(LR.1, LR.2)
BIC(LR.1, LR.2)

# Hosmer-Lemeshow goodness-of-fit test
library(ResourceSelection)
hoslem.test(healthy, fitted(LR.2), g = 10)
hlt2 <- hoslem.test(healthy, fitted(LR.2), g = 10)
cbind(hlt2$observed, hlt2$expected)

anova(LR.2, update(LR.2, ~1), test="Chisq")

# Model comparison using the log likelihood ratio test
anova(LR.1, LR.2, test = "Chisq")

# Marginal effects
library(margins)
marg2 <- margins(LR.2)
summary(marg2)

# Predicted probabilities
newdf <- data.frame(educ=c(12,14,16),
                    maritals=rep(mean(maritals), 3),
                    wrkfull=rep(mean(wrkfull), 3),
                    female=rep(mean(female), 3))
newdf
newdf[,c('pred.prob')] <- predict(LR.2, newdata=newdf, type="response")
newdf

```

```
# Predicted probabilities with ggpredict() in ggeffects
library(ggeffects)
margins <- ggpredict(LR.2, terms="educ[12, 14, 16]")
margins
as.data.frame(margins)
sqrt(diag(vcov(margins)))
plot(margins)

margins.ew <- ggpredict(LR.2, terms = c("educ[12, 14, 16]", "wrkfull"))
margins.ew
plot(margins.ew)

# Probit regression
PR.2 <- glm(healthy ~ wrkfull + maritals + female + educ, data = chp3.1r, family =
binomial(link = "probit"))
summary(PR.2)

# Making tables to display the results
library(stargazer)
stargazer(LR.1, LR.2, type="text", align=TRUE, out="lr2mod.txt")
stargazer(LR.1, LR.2, type="html", align=TRUE, out="lr2mod.htm")

detach(chp3.1r)
```

## Glossary

**A Bernoulli distribution** is a distribution for a binary variable with 1 for having an event or success and 0 for not having an event or failure. It is a special case of the binomial distribution with one trial for each individual.

**A logit coefficient** is the regression coefficient in the logit scale in logistic regression models.

**A marginal effect** is a change in a response variable related to the change in an independent variable. In logistic regression, it is a change in the predicted probability with a change in the values of a predictor variable.

**A probit regression model** is used with the probit transformation of a binary outcome variable.

**An odds ratio (OR)** is a ratio of two odds.

**Deviance** compares the currently fitted model and the saturated model. It is defined as  $-2$  times the difference in log likelihood between these two models.

**Generalized linear models** extend the linear regression model when the response variable follows a distribution from the exponential family. They include three common components: the random component, the systematic component, and the link function.

**Logistic regression** is a regression method used to predict a binary outcome variable with the logistic transformation.

**The AIC** penalizes the deviance by the number of predictors.

**The BIC** penalizes the deviance by its degrees of freedom and the sample size.

**The likelihood ratio test** is the difference in deviance between nested models. It is the difference in  $-2LL$ , which can be expressed as a ratio of likelihood in logarithm.

**The linear probability model** is used to estimate a binary outcome variable with the linear regression method.

**The odds** are the ratio of two probabilities, the probability of success or of having an event ( $p$ ) to the probability of failure or of not having an event ( $1 - p$ ).

## Exercises

Use the GSS 2016 data available at <https://edge.sagepub.com/liu1e> for the following problems.

1. Conduct a logistic regression analysis to examine the relationship between the outcome variable `gunlaw` and three predictor variables from the three predictor variables `sex`, `educ`, and `age`. Before conducting the analysis, recode `gunlaw` into a new variable named `gun` so that 1 = favoring gun permit and 0 = opposing gun permit.
2. Conduct the likelihood ratio test of the overall model and interpret it.



3. Compute the deviance statistic for the model.
4. List three measures of pseudo  $R^2$  and the AIC statistic.
5. Identify the logit coefficient, the Wald  $z$  test, and the 95% confidence interval for the predictor variable `educ`.
6. Compute the odds ratios for `educ` and `age`.
7. Make a publication-quality table containing the estimated logit coefficients.
8. Write a report to summarize the results from the output.