Chapter 12

Logistic Regression

with Amr ElRafey

## [H1] Learning Objectives

**[INSERT NL]**

1. Calculate a logistic regression
2. Interpret the findings of logistic regression
3. Test the assumptions of logistic regression
4. Create a logistic regression using the ordinary regression of probabilities
5. Create a propensity-to-seek-treatment score
6. Test hypotheses that changes in different contexts

**[END NL]**

## [H1] Key Concepts

**[INSERT BL]**

* Natural logarithm function
* Logistic regression
* Coefficients of regression equation
* Context dependence
* Chi-square goodness of fit
* Deviance
* Hosmer-Lemshow tests

**[END BL]**

## [H1] Chapter at a Glance

The distribution of the dependent variables dictates what type of regression is appropriate. If the dependent variable is continuous, multivariate regression works best. We reviewed this in chapter 11, which focused on ordinary multiple regression. When the dependent variable is a count of events, then Poisson regression is used (see chapter 19). When the outcome variable is binary (i.e., assumes the value of 0 or 1), logistic regression is appropriate. This chapter discusses logistic regression, showing that patients differ in their reasons for seeking treatment. In chapter 13, logistic regression is used to estimate propensity for participation in treatment, and the resulting propensity scores are used to control patient characteristics statistically.

[H1] Widespread Use

Logistic regression is used often in healthcare management. While the following discussion of the use of logistic regression is not exhaustive, it is sufficient to show where students in health administration might see these applications.

**[H2]** Management and Leadership Courses

Logistic regression can be used to identify factors that predict nurse retention; for example, Colville and colleagues (2017) examined burnout among staff in intensive care units. Nylinder (2009) used logistic regression to show that perception of a manager’s tight budgetary control depends on a number of factors, including the actual budget environment, the gender of the manager, and whether the manager was a physician or nonphysician. Vera and Hucke (2009) used logistic regression to predict the career success of physicians and the role of a physician’s management orientation (e.g., their approach to considering financial goals, how they exercise management skills).

**[H2]** Human Resources Courses

Kheirbek and colleagues (2016) used logistic regression to predict whether white and black patients would differ in health outcomes such as rehospitalization. They found that the two groups differed in their conditions on admission but not in the quality of care they received. Logistic regression can also be used to think through recruitment issues. Chen, Blumenthal, and Jena (2018) used logistic regression to examine what type of physicians were more likely to be excluded from participation in Medicare owing to fraud, waste, and abuse (result: male, older, with osteopathic training). Ariza-Montes and colleagues (2013) used logistic regression to identify workplace bullying and the factors that foster this behavior. They found that bullying increases among those who work on a shift schedule, perform monotonous and rotating tasks, suffer from work stress, enjoy little satisfaction from their working conditions, and do not perceive opportunities for promotions in their organizations.

**[H2]** Courses on Quality

Kodama and Kanda (2010) used logistic regression to learn how risk managers elect to address near-miss events. They found that priorities for addressing near-misses were not always based on the extent of harm to the patient. These priorities were also affected by how rare the event was, the effect of the event on the reputation of the nurse, and the possibility of delayed discharge. In program evaluation and quality courses, logistic regression may be used to see how various parts of an organization report events. Basu and Friedman (2013) used it to see whether Medicare health maintenance organizations reported more adverse outcomes than Medicare fee-for-service organizations. After adjusting for quality of care, fee-for-service organizations reported fewer iatrogenic pneumothoraxes, accidental punctures or lacerations, and postoperative respiratory failures.

**[H2]** Courses on Health Insurance, Health Policy, or Risk Assessment

Shin and colleagues (2010) used logistic regression to identify factors that contribute to the overuse of medical services. They identified that overusers were elderly, female, less educated, and dealing with stressors, and that they had lower perceived health status. Access to care and services are often analyzed using logistic regression. Robin and colleagues (2013) found that uninsured or publicly insured cancer survivors were less likely to have a usual source of care or access to preventive services.

**[H2]** Courses on Marketing

Logistic regression can be used to identify factors that contribute to increased market share. Eslami and colleagues (2009) examined the market share of intervention radiologists, vascular surgeons, and interventional cardiologists in the management of peripheral arterial interventions. Vascular surgeons had the best outcomes but not the largest market share.

Khalil Zadeh, Robertson, and Green (2017) report the effect of pharmaceutical advertisement on at-risk patients. They found that patients who were older, less educated, and lower income were more vulnerable to drug advertising, as well as ethnic minorities and patients with poor health status.

**[H2]** Courses on Strategy

Kim (2016) used logistic regression to find factors that affect five-star ratings in the Nursing Home Compare website. He found that nursing homes working in competitive markets, where there is excess supply, had worse reports.

Logistic regression is also often used to analyze hospital acquisitions. Noles and colleagues (2015) found that hospitals with weaker financial performance but lower staffing levels and staffing costs were more likely to merge or be acquired. After a merger, these hospitals experienced declines in operating margins and less salary expense.

As these examples show, managers use logistic regression in a wide variety of situations. Most healthcare management students will see logistic regression in one of the courses ahead of them.

## [H1] Case Study

In this chapter, we continue with the medical foster home (MFH) example that we first discussed in chapter 11. In that chapter, multivariate regression was used to predict cost of care—a continuous variable. In the current chapter, the dependent variable is participation in the MFH program, a binary variable. The independent variables include six-month survival, age, gender, and 249 covariates measuring the resident’s medical history prior to admission. Data were collected from 2008 to 2015. Residents without their year of birth, who were less than 40 (younger and older populations differ in their utilization), who had fewer than 365 days of follow‑up, or who had negative costs were excluded from the data. Medical history was organized into diagnostic categories following the Clinical Classification Software (CSS) of the Agency for Healthcare Research and Quality.

In chapter 11, we considered whether MFH programs save money. In this chapter, we ask whether specific subgroups of patients are more likely to seek placement in an MFH. Residents who select one of these facilities may be different from residents who remain in the nursing home. For example, community homes may refuse to accept residents who are disruptive. In addition, sick residents may be less likely to change from nursing home. Residents in an MFH may not be equivalent to those in a nursing home, unless we are comparing the same types of patients across the two programs. At a minimum, we would expect the residents in both programs to have the same medical history.

Data from electronic health records are observational. In general, when judging based on observational data, participants and nonparticipants in a program would differ in many characteristics. The possibility of confounding among the variables rules out a simple comparison of mean cost for program participants and nonparticipants. Chapter 12 introduces the concept of propensity scoring so that treated and untreated patients have the same patterns of covariates. Propensity scoring reduces differences between MFH and nursing home patients by weighting the cases so that the two groups have similar medical histories. In this chapter, we use logistic regression to create an index of patients’ propensity to participate in treatment, then use this index to remove confounding in observational data.

The first question is what types of patients choose an MFH or end up being admitted to the program. The response variable—the dependent variable—in this analysis is participation in MFH. Let us assume that nursing home residents either continue to live in the nursing home or choose to live in a community home. This response variable will be 1 when the patient resided in MFH and 0 when the patient resided in a nursing home. The second question to answer is how we can use the propensity to participate in the MFH to remove confounding in the data.

## [H1] Logistic Regression Model

We want to predict from the residents’ age, gender, and medical history their likelihood of participating in the MFH program. One possibility is to use ordinary regression to establish a model that relates the probability of participating in an MFH to independent variables. Such a regression might work well, but it would have predicted values that may exceed 1; thus, we could not interpret the predicted values as a probability estimate. By definition, a probability function must range between 0 and 1. Probabilities above 1 are meaningless. To make sure that the predictions are in the appropriate range, the analyst regresses the log of odds of participation on the predicted values. Log odds range from 0 to infinity, so they do not have the problem of out-of-range predictions. Once the log odds of participation have been calculated, the probability of the participation can be calculated by simple transformation. Let us suppose that (read out loud as “pie”) indicates the probability or propensity of the response variable; in the case of the MFH program, indicates the propensity of participating. The odds of participation in an MFH are given by the equation

**[INSERT EQUATION]**

.

**[END EQUATION]**

The response variable is then calculated as a natural logarithm of the odds of participation in the program, as

**[INSERT EQUATION]**

.

 **[END EQUATION]**

The logarithm function can be calculated for different bases, typically for base 10. The natural logarithm is calculated at a mathematical constant *e*, where *e* is 2.718281828459. The *e* index in highlights that the logarithm is calculated at base *e*; an easier way to display this is to drop *e* and change its abbreviation from “” to “ln,” emphasizing that it is a natural logarithm (so called because the constant *e* occurs often in nature). To get a sense of how natural logarithms transform the data, it is helpful to see the functions for different numbers (exhibit 12.1). The natural logarithm of a number increases slowly when numbers are larger than 1. It declines rapidly when numbers are below 1. It goes to negative infinity as the number gets closer to 0.

**[INSERT EXHIBIT]**

**Exhibit 12.1** Natural Logarithm Function

**[END EXHIBIT]**

The logistic regression for predicting from *n* independent variables,, is calculated as:

**[INSERT EQUATION]**

.

**[END EQUATION]**

The function is called the *logit function* of . Logit function is one of many transformations that allow analysts to calculate binary responses from a set of predictors. These transformations are also called *link functions*, in the sense that they link the independent variable to the transformed response variable. The analyst can always predict the probability of an event by using the inverse of the link function. For example, the probability of participation in the MFH is calculated in two steps. First, the logistic regression is used to predict the log of odds of participation. Then, second, the inverse of logit function is used to transform the predicted value to an estimate of the propensity to participate. Assume that the predicted value from the regression is (pronounced “LAM-duh”), written as

**[INSERT EQUATION]**

.

**[END EQUATION]**

If this is the case, the propensity of participation in the MFH program is calculated as

**[INSERT EQUATION]**

.

**[END EQUATION]**

Because is a probability function, the variance of the predicted values can be calculated as The procedure allows us not only to predict the probability of the event, but also to calculate its variance and thus to construct a confidence interval around our predictions.

## [H1] Example of Ordinary Regression with Logit Transformation

In this section, we show how logit transformation can be used to conduct ordinary regression with binary data. For this example, we use the MFH data. The dependent variable is participation in the program. The independent variable are the various disabilities lasting for 365 days. The following set of codes reads the data and substitutes values for missing or null values.

**[LIST FORMAT]**

Drop Table #Data

Declare @AvgAge as Float

Set @AvgAge = (SELECT avg(age) FROM [MFH].[dbo].[DailyCost$])

SELECT distinct [MFH]

 , iif([bathing\_365]>.5,1,0) AS Bathing

 ,iif([bladder\_365]>.5, 1,0) as Bladder

 ,iif([bowelincontinence\_365]>.5, 1,0) as Bowel

 ,iif([dressing\_365]>.5, 1, 0) as Dressing

 ,iif([eating\_365]>.5, 1, 0) as Eating

 ,iif([grooming\_365]>.5, 1, 0) as Grooming

 ,iif([toileting\_365]>.5, 1, 0) as Toileting

 ,iif([transferring\_365]>.5, 1, 0) as Transferring

 ,iif([walking\_365]>.5, 1, 0) as Walking

 ,iif([gender] ='F', 0, 1) AS Male

 ,iif([race] ='B', 1,0) as Black

,iif(race='W', 1, 0) AS White

,iif(race='NULL', 1,0) AS RaceNull

 ,iif(age is null, @avgAge, age) AS Age

 , ID

INTO #Data

FROM [MFH].[dbo].[DailyCost$]

-- (39139 row(s) affected)

**[END LIST]**

Next, we divide the data into groups and calculate the probability of joining the MFH program in each group.

**[LIST FORMAT]**

DROP TABLE #Prob

SELECT CAST(Sum(MFH) as Float)/Cast(Count(distinct ID) as float) AS Prob

, count(distinct id) as n

, Bathing, Bladder, Bowel, Dressing, Eating, Grooming, Toileting

, Transferring, Walking, Male, Black, White, RaceNull, Floor([age]/10)\*10 AS Decade

INTO #Prob

FROM #DATA

GROUP BY Bathing, Bladder, Bowel, Dressing, Eating, Grooming, Toileting

, Transferring, Walking, Male, Black, White, RaceNull, Floor([age]/10)\*10

Having Count(distinct ID)>9

-- (405 row(s) affected)

**[END LIST]**

We can now transform this probability to a logit function using the following code:

**[LIST FORMAT]**

SELECT CASE

WHEN Prob=0 THEN log(1/cast(n as FLOAT) )

WHEN Prob=1 Then Log(Cast(n as FLOAT)/(Cast(n as FLOAT)+1.))

ELSE log(Prob/(1-prob)) END AS Logit, \* FROM #Prob
**[END LIST]**

In the last step, we can regress the transformed values on the disabilities and demographic variables. We used the ordinary regression command inside Excel. You can see the result in exhibit 12.2.

**[INSERT EXHIBIT]**

**Exhibit 12.2** Predicting Odds of Joining the Medical Foster Home Program

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|

|  |
| --- |
| SUMMARY OUTPUT |
|  |  |
| Regression Statistics |
| Multiple R | 0.325 |
| R squared | 0.105 |
| Adjusted R squared | 0.073 |
| Standard error | 1.277 |
| Observations | 405 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  |  |  |
|  | df | SS | MS | F | Significance F |
| Regression | 14 | 74.91 | 5.35 | 3.28 | 0.00 |
| Residual | 390 | 636.43 | 1.63 |  |  |
| Total | 404 | 711.34 |   |   |   |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Coefficients | Standard Error | *t* Stat | *P*-value | Lower 95% | Upper 95% | Lower 95.0% | Upper 95.0% |
| Intercept | −1.977 | 0.562 | −3.517 | 0.000 | −3.083 | −0.872 | −3.083 | −0.872 |
| Bathing | −0.691 | 0.228 | −3.031 | 0.003 | −1.139 | −0.243 | −1.139 | −0.243 |
| Bladder | 0.341 | 0.178 | 1.917 | 0.056 | −0.009 | 0.691 | −0.009 | 0.691 |
| Bowel | 0.046 | 0.188 | 0.245 | 0.806 | −0.324 | 0.416 | −0.324 | 0.416 |
| Dressing | −0.239 | 0.186 | −1.285 | 0.200 | −0.605 | 0.127 | −0.605 | 0.127 |
| Eating | 0.097 | 0.153 | 0.637 | 0.525 | −0.203 | 0.398 | −0.203 | 0.398 |
| Grooming | 0.090 | 0.164 | 0.546 | 0.586 | −0.233 | 0.412 | −0.233 | 0.412 |
| Toileting | 0.070 | 0.168 | 0.418 | 0.676 | −0.260 | 0.401 | −0.260 | 0.401 |
| Transferring | 0.646 | 0.148 | 4.364 | 0.000 | 0.355 | 0.937 | 0.355 | 0.937 |
| Walking | −0.183 | 0.200 | −0.917 | 0.360 | −0.575 | 0.209 | −0.575 | 0.209 |
| Male | −0.230 | 0.326 | −0.707 | 0.480 | −0.871 | 0.410 | −0.871 | 0.410 |
| Black | 0.325 | 0.351 | 0.928 | 0.354 | −0.364 | 1.015 | −0.364 | 1.015 |
| White | 0.157 | 0.334 | 0.470 | 0.638 | −0.499 | 0.813 | −0.499 | 0.813 |
| Race null | −0.095 | 0.362 | −0.263 | 0.793 | −0.807 | 0.617 | −0.807 | 0.617 |

 |

**[END EXHIBIT]**

This regression output shows the relationship between the logarithm of the odds of joining a MFH program and various disabilities, several of which have a statistically significant relationship to the dependent variable. From these data, we can conclude that residents with bathing disability are less likely to join the program; residents with bladder incontinence or inability to rise to a sitting position (transfer problems) are more likely to join the program.

## [H1] Predictors of the Use of an MFH Using R

The following section shows the propensity of joining the MFH program as a function of a large number of variables, each indicating a particular diagnosis in the patient’s medical history. The predictors that are statistically significant indicate differences in the age, gender, and medical history of patients who stay at a MFH or a nursing home.

|  |
| --- |
| To read the data into R, we use the read.csv() function as **[LIST FORMAT]**.**[END LIST]**This command tells R to import data from the CSV file called “MFHdata.csv” in a data frame titled “mfh.” (A data frame is a collection of variables and their related names in a matrix format that could be easily analyzed.) Note that R is case-sensitive and capitalization matters. To perform a logistic regression we use the glm() function (generalized linear models). The R code for using logistic regression is given next. In this code, “mfh” is the name of the data file. The daily cost reported for the resident is in the first data column. Columns 2 through 264 are independent variables. The distribution family is binomial, indicating the need for logistic regression. The full call to the glm package is**[LIST FORMAT]**.**[END LIST]**To view the results of the regression, we use the summary() function as follows: **[LIST FORMAT]**.**[END LIST]**Because there are hundreds of independent variables, the output is long, and we truncate the list of the variables to the first 30 of the 263 variables. **[LIST FORMAT]****[END LIST]** |

The findings confirm that MFH residents differed in a large number of diagnostic abilities (e.g., CCS10 “Immunizations and screening for infectious disease,” CCS100 “Acute myocardial infarction”) and functional abilities (e.g., dressing, eating, grooming, transferring) from nursing home residents. Based on these results, as expected, it does not make sense to compare the cost of these two programs without adjusting for differences in the patient populations they serve.

## [H1] Interpretation of Coefficients

The logistic regression equation enables the analyst to calculate the probability that a binary event might occur. It also provides an estimate for the parameters in the regression equation. Given that we can also estimate the variance of the coefficients, we can use logistic regression to test the hypothesis that no relationship exists between a predictor and the response (or dependent variable). In particular, we can test the hypothesis that the estimated coefficient comes from a distribution with a mean of zero.

The estimation of the parameters of a logistic regression requires the computer to iteratively consider different parameters, finding a set that best fits the data. The procedure follows the principles of maximum likelihood estimation. Different parameters are tried, the fit to the data is calculated, and the parameter with the best fit is reported. When the sample size is large, the parameters that produce the maximum likelihood of observing the actual outcome are selected. This procedure leads to estimations that have little bias, provides estimates of the standard deviation of the parameters, and assumes the distribution of the parameter is approximately normal. Unfortunately, there are no equations to estimate the parameters that can be easily solved; the iterative maximum likelihood approach is the best way forward. With the modern computer, the estimation procedure can be carried out quickly.

The test of the parameters of the regression coefficients is known as the *Wald test*. The maximum likelihood estimate for a coefficient of a predictor, is shown with a hat over the coefficient, to emphasize that it is an estimate from the sample. This parameter, its standard deviation, and its standard error are estimated from the sample using the maximum likelihood estimation procedure. The null hypothesis that the estimated parameter is a particular value, say, can then be tested using normal distribution, like so:

**[INSERT EQUATION]**

.

**[END EQUATION]**

Most common is to test that the variable has no relationship with the response variable, in which case the hypothesized coefficient is zero, calculated as

**[INSERT EQUATION]**

.

**[END EQUATION]**

The *z*-statistic can be tested by comparing it to standard normal distribution (see *z* values in the output from the analysis of MFH data). In addition, a confidence interval can be constructed for each parameter based on a *z*-multiplier (e.g., 1.96 for a two-sided 97.5 percent confidence interval). Because the confidence interval is expressed in terms of the natural logarithm of odds, the inverse of the logarithm function (the antilogarithm) can be used to reexpress it in terms of the odds of the response variable.

 Let us look at some examples. In the MFH case study, we can see estimates of parameters for different predictors and the Wald test of its significance (shown in the table as *z*-values). We can take, for example, the variable *dressing disability*, which has a coefficient of −1.556. This coefficient means that residents in the MFH are less likely to have dressing disability. The standard error, as reported by the regression model of this variable shown earlier, is 0.4582. As such, a Wald statistic would be calculated by dividing the coefficient by the standard error, so . The logistic regression model is also reporting a *z*-value of −3.395.

## [H1] Context Dependent Hypothesis Test

Chapters 6 and 7 discussed methods of hypothesis testing. There we saw how analysts can test whether a sample mean comes from a distribution with a mean or a particular rate. We just argued that we can also do so in a regression equation. The logistic or the multivariate regressions provide alternative ways of doing the same test, but surprisingly, the results are different. This is alarming. Statisticians expect different methods to lead to the same conclusion. So it is important to understand why tests of coefficients in a regression will lead to a result that is different from simply testing the same hypothesis from the distribution of the variable.

When testing the coefficients of a predictor in regression, we are testing the hypothesis in the context of other variables and events. Multicollinearity among the predictors leads to each predictor affecting the coefficient of other predictors. In predicting participation in the MFH program, the coefficient for bowel incontinence is not statistically significant (see exhibit 12.2). If we look only at this variable, we might find that MFHs and nursing home programs could differ significantly. The test of the coefficient of the predictor in a model will commonly lead to conclusions that differ from the test of the distribution of the predictor outside the model.

This discrepancy raises a problem with interpreting contradictory findings. What can we conclude when two different tests of the same concept have different results? The obvious answer is that the right solution depends on how we ask the question. In regression, we are testing the hypothesis in the context of other variables. In a comparison of distributions, we are not. If context matters, and it often does, we should test the hypothesis using regression models.

In real life, we are almost always interested in testing a hypothesis within a particular context. For example, if we want to test whether teaching patients on discharge about their symptoms (e.g., by having the nurse call them on the phone) reduces readmission, we want to know if the context of other variables also affects readmissions (e.g., severity of the patient’s illness on admission, level of patients’ cognition, caregiver’s prior knowledge).

If they want to test the hypothesis in context, the analyst concentrates on the coefficient for patient education in a regression equation. However, if context doesn’t matter, hypotheses should be tested by comparing distributions. Managers almost always pose questions for which context matters.

## [H1] Measures of Goodness of Fit

A model’s goodness of fit should be evaluated in a validation data set that has been set aside. Typically, data are randomly divided into training and validation sets. Regression parameters are estimated in the training set, but the goodness of fit is reported in the validation set. This type of cross-validation allows statisticians to avoid overfitting a model to one data set. Overfitting is *modeling the noise in the data*. Overfitting occurs often—every data set has unique characteristics that affect findings but are not present in other data. It also occurs when the number of variables in the model is larger than the number of observations. Cross-validation ensures that we are not modeling the noise. It ensures that relationships that exist in one data set, but not the other, are ignored.

Several different statistics measure goodness of fit or the predictive accuracy of a model. These statistics include the following:

**[INSERT BL]**

* *Chi-square goodness of fit tests*. In logistic regression, the difference between the predicted value of an outcome and its observed value is referred to as the residual. Given that the standard deviation of the prediction for a binomial outcome is calculated as (1− ), a standardized residual and a chi-square test statistic for these standardized residuals are calculated as

**[INSERT EQUATION]**

**[END EQUATION]**

* *Deviance*. The natural logarithm of the ratio of observed and predicted values can also report the accuracy of predictions. The formulas used to do so include

**[INSERT EQUATION]**

**[END EQUATION]**

Note that when the outcome is present (i.e., when we would want to predict a probability close to 1. The further the prediction is from 1, the worse the score in the formula. When the reverse happens. Now, we want to have a low probability of the outcome. The lower the predicted value, the better the prediction. The combined effect of these two formulas is called *deviance* (*D*) and is calculated with the equation

**[INSERT EQUATION]**

 **[END EQUATION]**

In this combined formula, the larger the deviance, the worse the fit.

* *Predicted and actual outcomes.* These can be organized into a contingency table. Chapter 4 discusses these types of contingency tables. The columns are actual observed values, and the rows are predicted values. If predicted values exceed a cutoff, then we predict that the outcome will happen; otherwise, we predict that it will not. These tables can be analyzed to estimate the sensitivity and specificity of predictions at different cutoff points.

A receiver operating curve (ROC) will show the relationship between the calculated sensitivity and the specificity of the predictions of the logistic regression. Chapter 5 introduced the ROC curve, and in it, we provided structured query language (SQL) code for measuring the area under the ROC. This area is also known as the *c*‑statistic. A perfect prediction will have a *c*-statistic of 1, and a random prediction will have a *c*-statistic of 0.

* *Pseudo-R2*. This statistic is the ratio of the variance of predicted values divided by the variance of observed values. It is best to think of it as portion of the variance in observed value explained by the predictions. This statistic ranges from 0 to 1, with 1 indicating a perfect prediction.
* *Hosmer-Lemshow tests*. In this test, data are organized into strata. The cases are listed in order of predictions, and cases that are adjacent are combined until 10 or 20 cases fall in each stratum. Each stratum represents a particular combination of covariates. In each stratum, the difference of observed and predicted values provides an estimate of the residual. A chi-square test on the observed versus expected predictions within these strata provide the Hosmer-Lemshow test.

**[END BL]**

 Keep in mind that perfect prediction is not possible. There are practical limits to predictions. The future is inherently unknown. Some patients beat the odds with better-than-expected outcomes. Measurement errors and other unknown factors affect the outcome. Not everything in life can be anticipated. Also keep in mind that in large data sets, the ability to predict outcomes becomes increasingly harder as sources of variability increase. When millions of patients are involved, the analyst expects that predictive models will explain a relatively small portion of the variation in outcomes.

## [H1] Chapter Summary

This chapter introduced logistic regression, a widely used method of analysis. We showed how this method can be used to analyze any binary dependent variable. It can also be used to predict the probability of receiving a particular treatment (also known as a propensity score). Propensity scores indicate how patients’ characteristics affect participation in treatment programs.

## [H1] Supplemental Resources

A problem set, solutions to problems, multimedia presentations, SQL code, and other related material are in the course website.

**[H1] References**

Ariza-Montes A., N. M. Muniz, M. J. Montero-Simó, and R. A. Araque-Padilla. 2013. “Workplace Bullying Among Healthcare Workers.” *International Journal of Environmental Research and Public Health* 10 (8): 3121–39.

Basu, J., and B. Friedman. 2013. “Adverse Events for Hospitalized Medicare Patients: Is There a Difference Between HMO and FFS Enrollees?” *Social Work in Public Health* 28 (7): 639–51.

Chen, A., D. M. Blumenthal, and A. B. Jena. 2018. “Characteristics of Physicians Excluded from US Medicare and State Public Insurance Programs for Fraud, Health Crimes, or Unlawful Prescribing of Controlled Substances.” *JAMA Network Open* 1 (8): e185805.

Colville, G. A., J. G. Smith, J. Brierley, K. Citron, N. M. Nguru, P. D. Shaunak, O. Tam, and L. Perkins-Porras. 2017. “Coping With Staff Burnout and Work-Related Posttraumatic Stress in Intensive Care.” *Pediatric Critical Care Medicine* 18 (7): e267–73

Eslami, M. H., N. Csikesz, A. Schanzer, and L. M. Messina. 2009. “Peripheral Arterial Interventions: Trends in Market Share and Outcomes by Specialty, 1998–2005.” *Journal of Vascular Surgery* 50 (5): 1071–78.

Khalil Zadeh, N., K. Robertson, and J. A. Green. 2017. “‘At-risk’ Individuals’ Responses to Direct to Consumer Advertising of Prescription Drugs: A Nationally Representative Cross-sectional Study.” *BMJ Open* 7 (12).

Kheirbek, R. E., J. Wojtusiak, S. O. Vlaicu, and F. Alemi. 2016. “Lack of Evidence for Racial Disparity in 30-Day All-Cause Readmission Rate for Older US Veterans Hospitalized with Heart Failure.” *Quality Management in Health Care* 25 (4): 191–196.

Kim, A. S. 2016. “Market Conditions and Performance in the Nursing Home Compare Five-Star Rating.” *Journal of Health Politics, Policy, and Law* 41 (5): 939–68.

Kodama, S., and K. Kanda. 2010. “Nurse Risk Managers’ Criteria for Dealing with Near-Miss Events.” *BioScience Trends* (5): 231–38.

Noles, M. J., K. L. Reiter, J. Boortz-Marx, and G. Pink. 2015. “Rural Hospital Mergers and Acquisitions: Which Hospitals Are Being Acquired and How Are They Performing Afterward?” *Journal of Healthcare Management* 60 (6): 395–407.

Nylinder P. 2009. “Tight Budgetary Control: A Study of Clinical Department Managers’ Perceptions in Swedish Hospitals.” *Journal of Health Services Research and Policy* 14 (2):70–76.

Robin, Y. K, P. F. Short, S. Machlin, E. Dowling, H. Rozjabek, C. Li, T. McNeel, D. U. Ekwueme, and K. S. Virgo. 2013. “Access to Preventive Health Care for Cancer Survivors.” *American Journal of Preventive Medicine* 45 (3): 304–12.

Shin, S. M., M. J. Kim, E. S. Kim, H. W. Lee, C. G. Park, and H. K. Kim. 2010. “Medical Aid Service Overuse Assessed by Case Managers in Korea.” *Journal of Advanced Nursing* 66 (10): 2257–65.

Vera, A., and D. Hucke. 2009. “Managerial Orientation and Career Success of Physicians in Hospitals.” *Journal of Health Organization and Management* 23 (1): 70–84.