Logistic Regression

In: Learning Statistics Using R

By: Randall E. Schumacker

Pub. Date: 2017
Access Date: August 25, 2021
Publishing Company: SAGE Publications, Inc.
City: 55 City Road
Print ISBN: 9781452286297
Online ISBN: 9781506300160
DOI: https://dx.doi.org/10.4135/9781506300160
Print pages: 448-480

Logistic Regression

Introduction

The logistic function was invented in the 19th century to describe the growth of populations as well as the course of autocatalytic chemical reactions (Cramer, 2003). In the case of population growth, the model investigated the time and growth rate of populations. Early scholars soon realized that an exponential function used in logistic regression had value but also had certain limitations. Verhulst (1845) formalized his earlier work and coined the name *logistic function*, which today takes the form of the logistic equation. Pearl and Reed (1920) knew of the autocatalytic reaction curves, but not knowing of Verhulst's work, they independently derived the logistic curve function, which they fitted to U.S. census data and that of many other living populations (human and animal).

Logistics regression models are different from the previous ordinary least squares (OLS) multiple regression models (Kleinbaum, 1994). The previous multiple regression models predicted a dependent variable that was measured at the interval or ratio level, and the difference between the predicted value and the actual value yielded an error of prediction. A logistic regression model has a dependent variable that is dichotomous, having only 0 and 1 as coded values. In a logistic regression equation, we predict the probability of *Y*, for example, between 0 (*not admitted to a program*) and 1 (*admitted to a program*). The probability of a *Y* value between 0 and 1 for an individual is estimated by knowledge of their values on the independent predictor variables.

In our previous multiple regression models, the estimation method to compute the regression weights was called the *least squares criterion.* The regression weights were computed based on minimizing the sum of squared errors (i.e., the difference between the actual *Y* and predicted *Y* values). In logistic regression models, the regression weights are computed based on *maximum likelihood estimation*, which is an iterative estimation process that finds the optimum set of regression weights. There are other differences between the previous OLS regression and logistic regression methods. Logistic regression models have different overall model fit criteria, tests of regression weight significance, odds ratios, and effect sizes. The logistic regression results are therefore different from OLS regression results, yet our goal of prediction and explanation remains the same.

Assumptions

Logistic regression models have a different set of assumptions from OLS regression, the most notable being no assumption of normality. Logistic regression assumptions relate to the following:

1. There is no assumption of linear relationship between the dependent and independent

variables.

2. Dependent variable values do not need to be normally distributed.
3. There is no homogeneity of variance assumption among or within categories.
4. Normally distributed error terms are not assumed.
5. Independent variables need not be measured on the interval scale.
6. There are no missing data in the contingency table (nonzero cell counts).

Recall that OLS multiple regression required a normal distribution of the dependent variable, residuals with constant variance and independent of each other, residual errors normally distributed with an average zero value, and independent variables that were not *collinear*—that is, they were highly correlated. However, multicollinearity, linearity, independence of errors, range of *X* values, and the order of predictor variable entry in the logistic regression equation are important issues to consider, as well as the data conditions a researcher might face when using logistic regression. These are briefly explained.

Multicollinearity is the level of intercorrelation among the independent variables when predicting a dependent variable. It is an important concern in any type of regression model. If the independent variables explain more of their shared variance due to high intercorrelation, then predicting the variation in *Y* scores is reduced. The ideal regression equation is when all of the independent variables are correlated with *Y* but are not correlated among themselves. We know this would rarely happen in practice, so detecting its presence in a regression equation is important. The Pearson, part, and partial correlations are one method to explore predictor variable relations; another is to find software that computes a collinearity statistic in logistic regression. Violation of this assumption can lead to unstable predictor variable regression weights and standard errors, or an overall model fit criterion that is significant while the individual regression weights are not.

Linearity is the assumption that the dependent variable has a linear relation with the independent predictor variables. In logistic regression, the dependent variable is binary or dichotomous, so the assumption of linearity is greatly reduced. The predicted value for *Y* is called a *logit*, so the linearity assumption is between the logit values of the dependent variable and the continuous predictor variables. Hosmer and Lemeshow (2000) suggest ways to detect whether the linearity assumption is violated, for example, by conducting a Box-Tidwell test. The logistic regression model would include the predictor variables and their interaction terms [$X_i * \ln(X_i)$]. Nonlinearity would be indicated by statistically significant interaction terms. Violation of the linearity assumption, which is present with significant interaction terms, can lead to unstable or biased parameter estimates or predicted logit values of *Y* not increasing across values of *X* predictor variables.

Independence of errors is the assumption that the errors of prediction are not correlated. Violation of this assumption would lead to inflated tests of regression weights due to underestimated or smaller standard errors. Recall that *t* = Regression weight/Standard error. Therefore, a smaller standard error would yield larger *t*-test values, increasing the chance of them being statistically significant, which inflates the Type I error rates.

The *range of X* values is fixed—that is, a logistic regression model is only valid for those *X* values used in the analysis. The use of *X* values for predictor variables outside the range of those used to compute the

regression weights would provide prediction errors caused by biased predictor variable slopes and intercept. A related data condition that produces biased regression weights and standard errors is when the number of predictor variables and the sample size are equal. For example, a restriction of range occurs with three predictor variables and $N = 3$ subjects because it does not allow for estimation of regression weights.

Missing data in logistic regression are also important to understand. For example, a nominal predictor variable should not have a single coded value, that is, 1 = *college*, which would give a *nonzero cell count* for 0 = *not college.* There should be a reasonable number of 0 = *not college* and 1 = *college* values, otherwise estimation of the regression weight would be incorrect. Basically, logistic regression estimation is affected by a lopsided percent category allocation. Similarly, if the dependent variable is perfectly predicted, the model estimation is suspect due to *nonseparation* of the data.

Outliers affect the maximum likelihood estimation of regression weights (Croux, Flandre, & Haesbroeck, 2002). There are residual plots and diagnostic tests that help identify influential data points when outliers are present in the data. Basically, an influential data point is indicated by standardized residual values >3.0, *DfBeta* values >1, or leverage values $>(m+1)/N$, where m = number of independent variables (close to 0 = *little influence* and close to 1 = *large influence*).

The *sample size* also affects logistic regression because maximum likelihood estimation of the regression weights generally requires larger sample sizes (Long, 1997).

The *order of predictor variable entry* affects the estimation of regression weights (Schumacker, 2005; Schumacker, Anderson, & Ashby 1999; Schumacker, Mount, & Monahan, 2002). The authors discovered that the various selection criteria (L^2, z, log-odds ratio, R_L^2 model variance, and ΔC^2) for model fit provided different results in the presence of a different order of predictor variable entry. The use of all possible subset selection and ΔC^2 fit criteria was recommended to determine the best set of categorical independent predictor variables in logistic regression.

Logistic Regression Equation

The logistic regression equation is expressed similarly as the OLS multiple regression equation, except that the data values and estimation of the regression weights are different. We are still interested in predicting a Y_i value for each individual given knowledge of the X predictor variables. The equation has an intercept a (the value of Y when all predictor values are 0), regression coefficients b_j, and residual error e_i:

$$Y_i = a + b_1X_1 + \cdots + b_jX_j + e_i.$$

A heuristic data set with Y (1 = *admit*, 0 = *not admit*), X_1 (1 = *U.S. student*, 0 = *foreign student*), and X_2 (1 = *male*, 0 = *female*) would be coded as seen here, in the left column:

Y	X_1	X_2

1	1	1
1	1	0
1	0	1
1	0	0
⋮	⋮	⋮
⋮	⋮	⋮
0	1	1
0	1	0
0	0	1
0	0	0

Y	X_1	X_2
1	1	4
1	1	7
1	0	3
1	0	4
⋮	⋮	⋮
⋮	⋮	⋮
0	1	3
0	1	2
0	0	1
0	0	2

We would run a *binary* logistic regression analysis since the independent predictor variables are dichotomous (binary). We can also run logistic regression analyses when the independent predictor variables are binary and continuous; for example, X_2 now represents the number of support letters for admission. The data with a nominal and a continuous independent predictor variable are as as seen here, in the right column:

In logistic regression analysis, the dependent variable value is predicted as a logit value (natural log of the odds of the dependent variable) using the independent predictor variable regression weights estimated by maximum likelihood estimation. The logistic regression equation computes a probability value—the probability that the dependent variable value will occur between the values of 0 and 1. The log odds or risk ratio provides a way to interpret these predicted *Y* probability values.

Probability, Odds, Log Odds, and Logit

Menard (2000a) explains the concepts of probability, logit, and log odds/risk ratio when interpreting the predicted *Y* values in logistic regression. Each concept is important to understand when interpreting the results from logistic regression, whether one uses nominal or continuous predictor variables. I briefly explain each here and use these terms later in an applied example.

Probability

In logistic regression, the *Y* variable is scaled 0 and 1, a nominal level of measurement. The logistic regression equation computes a probability value for each individual between 0 and 1, depending on their values for the independent predictor variables. Therefore, $P(1) = 1 - P(0)$ and $P(0) = 1 - P(1)$, so that $P(1) + P(0) = 100\%$. The predicted probability values permit an interpretation of the *Y* outcome that is not binary (admitted/not admitted); rather, it is a matter of probability of occurrence—for example, “Do U.S. male students have a higher probability of admittance to a program than U.S. female students?”

The logistic regression equation also provides a classification prediction that permits a cross-tabulation with the actual *Y* binary values. A percent correct classification can be computed, which provides a helpful interpretation of the *Y* predicted outcome. A chi-square test would determine if the percent classification for actual and predicted values was statistically significant.

Odds and Logits

Odds are the ratio of the two probability outcomes for *Y.* The odds that the *Y* outcome is 1 (*admit*) rather than 0 (*not admit*) is the ratio of the odds (probability) that $Y = 1$ to the odds (probability) that *Y* is not equal to 1—that is, $1 - P(1)$. This can be expressed as

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

A few *odds* are expressed in the table between 0 and 1 as follows:

P(*1*)	*1* - P(*1*)	Odds ($P(1)/1 - P(1)$)
.001	.999	0.001
.250	.750	0.333
.500	.500	1.000
.750	.250	3.000
.999	.001	999.000

The *odds* do not provide the property of scale required to make meaningful interpretations. Therefore, the log of the *odds* is taken to create a scale that could range from positive to negative infinity. The *log odds* of *Y* creates a linear relationship between the probability of *Y* and the *X* predictor variables (Pampel, 2000). The log odds are computed by using the *ln* function on a scientific calculator. The log function converts the *odds* into positive and negative values, above and below the center point of 0, which is the value for .5/.5 = ln(1.000) = 0. The tabled values for the *odds* and *log odds*, ln(Odds) are seen here, to the left:

Odds	ln(Odds)
0.001	-6.907
0.333	-1.099
1.000	0
3.000	1.099
999.000	6.907

The logit values for *Y* are the log odds values—that is, Y_{logit} = ln(Odds). The logit values are related to the probabilities; that is, a one-unit change in the logit value equals a change in the probability. The logarithm function creates a linear scale, which permits a regression equation to predict the *log odds,* Y_{logit}:

$$\hat{Y}_{\text{logit}} = a + \beta_1 X_1 + B_2 X_2.$$

The logistic regression equation computes a predicted probability value for *Y.* The residual errors (*e*) are the differences between the actual *Y* values (0 or 1) and the predicted probability values.

The *log odds,* Y_{logit}, are not by themselves easy to interpret in relationship to the independent predictor variables. Therefore, to interpret each independent variable effect on the dependent variable, the exponent is taken, which converts back to an odds interpretation. The logistic regression equation is now expressed as follows:

$$\hat{e}^{Y_{\text{logit}}} = \left(e^{a}\right)\left(e^{\beta_1 X_1}\right)\left(e^{\beta_2 X_2}\right).$$

The logistic regression equation is now a multiplicative function due to the use of exponential values. On a scientific calculator, the symbol e^x is used for obtaining the exponential value. In the multiplicative equation, a value of $x = 0$ corresponds to $e^0 = 1$, which is a coefficient of 1. *Coefficients greater than 1 increase the odds, and coefficients less than 1 decrease the odds.* The log odds are linked to probabilities by the following equation:

$$P(1) = \frac{e^{a + \beta_1 X_1 + \beta_2 X_2}}{1 + e^{a + \beta_1 X_1 + \beta_2 X_2}}.$$

This linkage shows that the probability values close to 1 indicate more likelihood of the outcome—that is,

of being admitted to a program. Probabilities close to 0 would indicate a likelihood of not being admitted to the program. The probability is equal to the constant *e* or natural log function value, *e* = 2.718, with the exponent being the logistic regression equation. I showed that a probability ratio of .50/. 50 = 1 (odds ratio), with ln(Odds) = 0. The logit probability can be converted back, which equals a probability of .50:

$$P(1) = \frac{2.718^0}{1 + 2.718^0} = .50.$$

The logistic regression output provides the log odds value for the independent variables in the regression equation.

Odds Ratio versus Risk Ratio

The *odds ratio* (OR) is a ratio of the odds of an event, so it is not a probability value.

The *risk ratio* (RR) reflects the probability of an outcome in a group divided by the group's combined probability of success and failure, divided by a second outcome's probability in the group divided by that group's combined probability of success and failure. A risk ratio describes how many more times it is likely for an event to occur given the outcome. For example, RR = 3 for a treatment implies that one is three times more likely to achieve success with the treatment than without it. The following 2 × 2 table shows the cell values to compare the odds ratio and risk ratio computations:

The odds ratio is as follows:

$$\text{OR} = \frac{a/b}{c/d}.$$

	Success	Failure
Admit	*a*	*b*
Not admit	*c*	*d*

The risk ratio for admittance to a program is shown in the following formula:

$$\text{RR} = \frac{a/(a+b)}{c/(c+d)}.$$

An odds ratio is a measure of the association between a treatment and an outcome. The odds ratio represents the odds that an outcome will occur given the treatment, compared with the odds of the outcome occurring without the treatment. Case–control studies generally report an odds ratio due to predictor variable regression weights in a logistic regression equation. A logistic regression coefficient is estimated and then interpreted as the increase in the log odds of the outcome per unit increase in the value for the treatment group. The exponential function e^b is the odds ratio associated with a one-unit increase in the value for a predictor variable. The binary outcomes, or counts of success versus failure in a number of trials, are interpreted as an

odds ratio. The odds ratio can be interpreted as follows:

- OR = 1: The treatment does not affect the odds of the outcome.
- OR > 1: The treatment is associated with higher odds of the outcome.
- OR < 1: The treatment is associated with lower odds of the outcome.

A risk ratio describes how many more times it is likely for an event to occur given the outcome. Randomized control and cohort studies typically report relative risk using Poisson regression methods. The count of events per unit exposure have relative risk interpretations—that is, the estimated effect of a predictor variable is multiplicative on the rate, thus yielding a relative risk (risk ratio). The interpretation of a relative risk between an experimental group and a control group is as follows:

- RR = 1: There is no difference in risk between the two groups.
- RR > 1: The outcome is more likely to occur in the experimental group than in the control group.
- RR < 1: The outcome is less likely to occur in the experimental group than in the control group.

When the 2 × 2 table results have small cell numbers, *a* and *c* will be small numbers, so the odds and risk ratios will be similar. It is recommended that an odds ratio be converted to a risk ratio for interpretation, which can be done using the R *orsk* package (Wang, 2013). The risk ratio implies how many more times one is likely to achieve success (outcome) with a treatment than without it.

Model Fit

The model fit criteria are used to determine the overall fit of the predictor variables in predicting Y. For each unit change in the independent variable, the logistic regression coefficient represents the change in the predicted *log odds* value, Y_{logit}. The regression weights can also be tested for the statistical significance of each predictor variable. A final interpretation is possible using classification accuracy from the predicted group membership values. The first two, overall model fit and statistical significance of regression weights, are common when running regression-type analyses. The third, classification accuracy, is unique to logistic regression, and in the binary dependent variable, it affords a chi-square test of independence between the actual and predicted values.

Schumacker et al. (1999) also investigated the various model fit criteria used in logistic regression. Rather than compare *R*-squared analog values, the authors chose to examine many different selection criteria: L^2, *z*, log odds ratio, R_L^2, and ΔC^2. They discovered that ΔC^2 was a more robust measure of model fit and that the results differed depending on the order of predictor variable entry. I represent ΔC^2 as $\Delta\chi^2$ in this chapter. Menard (2000b) compared the different *R*-squared analog coefficients of determination in multiple logistic regression analysis. He recommended R_L^2 over the other *R*-squared analogs that were compared, based on its conceptual similarity to the OLS coefficient of determination and its relative independence from

the base rate. It seems prudent when conducting a logistic regression analysis to examine the overall model fit criteria as well as interpret the statistical significance of the predictor variables. Forward, backward, and stepwise methods for variable selection are not recommended because they lead to erroneous fit statistics, bias standard errors, and ill-determined predictor variable selection. An all-possible subset variable selection method would help determine the best set of predictor variables.

Chi-Square Difference Test

The test of overall model fit in logistic regression can use the likelihood ratio test. A log-likelihood (LL) function provides an index of how much has *not* been explained in the logistic regression equation after the parameters (regression weights) have been estimated. The LL values vary from 0 to negative infinity, with LL values close to 0 indicating a better logistic equation model fit. If we calculate LL for different logistic regression equations, then a test of the difference in LL indicates changes in the overall model fit. For example, calculate LL for a logistic regression equation using the intercept only. Next, calculate LL for a logistic regression equation with the intercept $+B_1X_1$. A difference in the LL values between these two regression models will indicate whether X_1 is statistically significant above and beyond the intercept-only model. This test is similar to the *F* test in multiple regression when testing a difference between the full and the restricted model; however, a chi-square difference test is used with the separate logistic regression chi-square values. A chi-square test is possible by multiplying by −2 the difference in the LL functions, with the degrees of freedom equal to the difference in the model degree-of-freedom values. The chi-square difference test is expressed as follows:

$$\Delta\chi^2 = -2\left(LL_{\text{full model}} - LL_{\text{restricted model}}\right)$$

and

$$df_{\text{Diff}} = df_{\text{full model}} - df_{\text{restricted model}}.$$

The larger the difference between the full and the restricted model, the better the model fit. Typically, a researcher starts with a baseline model (intercept only) and then sequentially adds predictor variables; each time, the degree-of-freedom difference would be 1. Each variable added in the logistic regression equation is tested for statistical significance given that $\chi^2 = 3.84$, $df = 1$, and $p = .05$. Consequently, any chi-square difference ($\Delta\chi^2$) greater than 3.84 would indicate a statistically significant predictor variable.

Note: The chi-square difference test is based on nested models—that is, adding additional variables from the same data set.

Hosmer-Lemeshow Goodness-of-Fit Test

The *Hosmer-Lemeshow goodness-of-fit* test is based on dividing the individuals into 10 groups (deciles) based on their predicted probabilities. A chi-square is computed for the observed and expected frequencies

in the 10 groups. A researcher would desire a nonsignificant chi-square statistic, which indicates that the predicted values are not different from the observed values given the logistic regression equation. The test is considered to lack power to detect the lack of model fit in the presence of nonlinearity of the independent variables. The test also tends to overestimate model fit when the groupings are less than 10 and provides little guidance about predictor variable significance. I mentioned it only because the test is often reported in statistical packages.

R-Squared Analog Tests

The *R*-squared analog tests are considered pseudo *R*-squared values or analogs to the *R*-squared value in OLS regression. As noted before, they have been investigated, with scholars disagreeing on their usefulness. The ones listed here are all variants when using the LL function.

Cox and Snell (1989) proposed the following with sample size = *n*:

$$R^2_{CS} = 1 - \exp\left(\frac{-2\text{LL}_{\text{model}} - \left(-2\text{LL}_{\text{baseline}}\right)}{n}\right).$$

Nagelkerke (1991) proposed an adjustment to the R_{CS}^2 value to achieve a maximum value of 1:

$$R^2_N = \frac{R^2_{CS}}{1 - \exp\left(-\frac{-2\text{LL}_{\text{baseline}}}{n}\right)}.$$

Hosmer and Lemeshow (2000) indicated an *R*-squared value for a ratio of the full model to the restricted model, thus indicating how the model can be improved by adding predictor variables:

$$R^2_L = \frac{-2LL_{\text{model}}}{-2\text{LL}_{\text{baseline}}}.$$

Harrell (2001) proposed an adjustment to the Hosmer and Lemeshow *R*-squared value for the number of independent variables (*m*) in the logistic regression equation:

$$R^2_{LA} = \frac{\left(-2LL_{\text{model}}\right) - 2m}{-2\text{LL}_{\text{baseline}}}.$$

The list goes on. The creation of *R*-squared analog tests used the same logic with the *R*-squared and *R*-squared adjusted values for sample size and number of predictor variables in OLS multiple regression. In fact, a researcher can compute an *R*-squared value by squaring the correlation between the observed *Y* values and the predicted *Y* probability values from the logistic regression equation. This would basically be as follows:

```
> Rsq = cor(Y, Yhat)^2
```

I will compute this *R*-squared value in the applied example given later in the chapter.

Predicted Group Membership

A final overall model fit approach can involve using the actual *Y* values (0, 1) with the predicted group membership. For example, if $P \geq .5$, then classify it as group = 1, else group = 0.

A crosstab table will provide the frequency and percentage of cases for the actual versus predicted probabilities for the groups. A correct classification would be for those individuals in the actual $Y = 1$ and predicted $Y = 1$ group and for those individuals in the actual $Y = 0$ and predicted $Y = 0$ group. The other two cell frequencies represent a misclassification. Therefore, the sum of the diagonal percents indicates the percent correctly classified, shown in the table below as percent correctly classified = $\sum(a + d)$.

Actual	Predicted	
	Admit	Not Admit
Admit	*a*	*B*
Not admit	*c*	*D*

The traditional chi-square test can be conducted using these actual and predicted values.

The Pearson chi-square formula with observed (*O*) and expected (*E*) values and $df = 1$ is

$$\chi^2 = \Sigma\frac{(O - E)^2}{E}.$$

An alternative form of the chi-square test of classification accuracy given a binomial outcome is as follows:

$$\chi^2 = \frac{\left[N_i - (np_i)\right]^2}{np_i},$$

where *N* = total sample size, *n* = number of cases in the cell, and *p* = probability of success. This provides the same Pearson chi-square value as the equation above.

Logistic Regression Coefficients

Maximum likelihood estimation is used to determine the logistic regression weights. The iterative estimation method starts with least squares estimates, and then maximizes the LL function for a final set of regression coefficients. The Wald and BIC (Bayesian Information Criterion) statistics are used to test the regression coefficients for statistical significance; standardized coefficients are computed separately, and a confidence interval is generally computed. Each of these is explained next.

Wald Chi-Square Test

The Wald chi-square test is computed as the logistic regression coefficient squared and divided by the square of the standard error (variance). This formula is shown below:

$$W = \frac{\beta_j^2}{\left(SE_{\beta_j}\right)^2}.$$

When conducting logistic regression, our approach is to first compute a baseline model (intercept only). We then add independent predictor variables, usually one at a time, and check to see if the Wald chi-square value is statistically significant for the regression coefficient.

AIC and BIC Tests

The other tests for the statistical significance of a regression coefficient are the AIC (Akaike Information Criterion) and BIC tests. The difference between the two measures is seen in the added component of either 2 * *df* for AIC, where *df is* the degrees of freedom associated with the LL function, or log(*n*) for BIC, where *n* is the sample size.

$$\text{AIC} = -2 * \text{LL} + (2 * df)$$

and

$$\text{BIC} = -2 * \text{LL} + \log(n).$$

The AIC and BIC values should be positive. The larger the AIC and BIC values, the more likely the added predictor variable is statistically significant; hence, it adds to the prediction in the logistic regression equation.

Standardized Coefficient

The beta regression weights (standardized coefficients) were used in OLS regression to assess which predictor variables were the most important and contributed to the prediction of *Y.*

In OLS regression, we obtained the standardized coefficients by first scaling the variables using the *scale()* function. This computed the standardized variables, which were then used in the regression function to output standardized beta weights. This is not done in logistic regression: rather, a test of the chi-square difference between a full and a restricted model is computed. In logistic regression, this involves using the $\Delta\chi^2$ test for the difference in models, where one variable is dropped to test whether it contributed significantly to prediction.

Confidence Intervals

A confidence interval (CI) around the logistic regression coefficient b_j provides important information for interpreting the coefficient. The confidence interval is formed using the regression coefficient plus or minus the product of a tabled critical value and standard error. The formula is expressed as follows:

$$\mathrm{CI}(b_j) = b_j \pm t_{\alpha/2;\, n-m-1}(S_{b_j}).$$

If the confidence interval for the regression coefficient contains zero, then the logistic regression coefficient is *not* statistically significant at the specified level of significance (α). We interpret confidence intervals to indicate how much the value might vary on repeated sampling of data.

Effect Size Measures

The *R*-squared analog coefficients are used as an effect size measure. They indicate the pseudovariance explained in predicted probabilities of *Y* given a set of independent predictor variables. The *odds ratio* is also an effect size measure, similar to the *R*-squared analog coefficients. The odds ratio is computed by the exponent of the regression weight for a predictor variable, e^{bj}, which is the odds for one group (admit) divided by the odds of the other group (not admit). When OR = 1, there is no relation between the dependent variable and the independent predictor variable. If the independent predictor variable is continuous, the odds ratio is the amount of change for a one-unit increase in the independent variable. If OR > 1, the independent variable increases the odds of the outcome, and when OR < 1, the independent variable decreases the odds of the outcome. The odds ratio can also be converted to a *Cohen's d* effect size:

$$\text{Cohen's } d = \frac{\ln(\mathrm{OR})}{1.81}.$$

Applied Example

The *aod* package is used to obtain the results for a logistic regression model analysis. It provides the functions we need to analyze counts or proportions. The *ggplot* package will be used to graph the predicted probability values. The *Hmisc* package will be used to obtain a data set for the logistic regression analysis. The packages are installed and loaded with the following R commands:

```
> install.packages("aod")
> install.packages("ggplot2")
> install.packages("Hmisc")
> library(aod)
> library(ggplot2)
> library(Hmisc)
```

I found several data sets in the Hmisc package:

```
> getHdata()
[1] "abm" "acath" "ari" "ari_other"
[5] "birth.estriol" "boston" "cdystonia" "counties"
[9] "diabetes" "dmd" "DominicanHTN" "FEV"
[13] "hospital" "kprats" "lead" "nhgh"
[17] "olympics.1996" "pbc" "plasma" "prostate"
[21] "rhc" "sex.age.response" "stressEcho" "support2"
[25] "support" "titanic2" "titanic3" "titanic"
[29] "valung" "vlbw"
```

I chose the data set *diabetes.* The data set appeared in the RGui window, and two separate Internet HTML windows opened with the documentation. The first HTML window indicated that the data set contained 19 variables on 403 of the 1,046 subjects who were interviewed in a study to examine the prevalence of obesity, diabetes, and cardiovascular risk factors for African Americans in central Virginia (Willems, Saunders, Hunt, & Schorling, 1997). The other HTML window provided a codebook for the variables in the data set. The R command was as follows:

Selecting Variables

```
> getHdata(diabetes, "all")
```

I attached the data set and listed the variable names to select the ones I wanted to use in the logistic regression equation.

```
> attach(diabetes)
> names(diabetes)
[1] "id" "chol" "stab.glu" "hdl" "ratio" "glyhb"
[7] "location" "age" "gender" "height" "weight" "frame"
[13] "bp.1s" "bp.1d" "bp.2s" "bp.2d" "waist" "hip"
[19] "time.ppn"
```

After examining the list of variable names, I selected *glyhb* (glycosolated hemoglobin), because levels >7.0 is a positive diagnosis of diabetes; *location* (Buckingham, Louisa); *gender (male*, female); and *frame* (small, medium, large) for my logistic regression analysis. I created a smaller data set, *viewdata*, with just these variables in them (*myvars*). I then used the *head()* function to print out the first 6 observations and *dim()* function to indicate the number of observations and number of variables (403 observations and 19 variables) in the data set. The following R commands were used:

```
> myvars = c("glyhb","location","gender","frame")
> newdata = diabetes[myvars]
> head(newdata, n = 6)
  glyhb location  gender frame
1 4.31 Buckingham female medium
2 4.44 Buckingham female large
3 4.64 Buckingham female large
4 4.63 Buckingham  male  large
5 7.72 Buckingham  male  medium
6 4.81 Buckingham  male  large

> dim(diabetes)
[1] 403 19
```

Handling of Missing Data

The first step prior to data analysis is to determine if missing values exist for one or more variables. The variables *glyhb* (*n* = 390) and *frame* (*n* = 391) had fewer than the original *N* = 403 observations. I used the *describe()* function to reveal information about each variable.

```
> describe (newdata)

newdata

4 Variables 403 Observations
---------------------------------------------------------------------------
glyhb: Glycosolated Hemoglobin
n missing unique Mean .05 .10 .25 .50 .75 .90
390 13 239 5.59 3.750 4.008 4.380 4.840 5.600 8.846
.95
10.916

lowest: 2.68 2.73 2.85 3.03 3.33, highest: 13.70 14.31 14.94 15.52 16.11
---------------------------------------------------------------------------
location
n missing unique
403 0 2

Buckingham (200, 50%), Louisa (203, 50%)
---------------------------------------------------------------------------
gender
n missing unique
403 0 2

male (169, 42%), female (234, 58%)
---------------------------------------------------------------------------
frame
n missing unique
391 12 3

small (104, 27%), medium (184, 47%), large (103, 26%)
```

We can delete missing values from our statistical analysis using *na.rm* = *TRUE*, or delete them from the data set prior to conducting statistical analysis. (*Note:* Alternative methods are used to impute missing data, which are beyond the scope of this book.) The data set with listwise deletions resulted in *N* = 379 subjects, a decrease of 24 subjects. The following R commands omitted the 24 subjects and then described the data:

```
> mydata = na.omit(newdata)
> describe(mydata)
mydata

4 Variables 379 Observations
---------------------------------------------------------------------------
glyhb: Glycosolated Hemoglobin
n missing unique Mean .05 .10 .25 .50 .75 .90
379 0 235 5.601 3.745 4.010 4.380 4.840 5.615 9.172
.95
10.934

lowest: 2.68 2.73 2.85 3.03 3.33, highest: 13.70 14.31 14.94 15.52
16.11
---------------------------------------------------------------------------
location
n missing unique
379 0 2

Buckingham (182, 48%), Louisa (197, 52%)
---------------------------------------------------------------------------
gender
n missing unique
379 0 2

male (158, 42%), female (221, 58%)
---------------------------------------------------------------------------
frame
n missing unique
379 0 3

small (102, 27%), medium (178, 47%), large (99, 26%)
---------------------------------------------------------------------------
> dim(mydata)
[1] 379 4
```

The following command lists the *structure* of the variables in the data set—that is, it shows the class and level of the variables in the data set, *mydata.* The variable *glyhb* is numeric (atomic), while *location* and *gender* are factors with two levels, and *frame* is a factor with three levels.

```
< str(mydata)
'data.frame': 379 obs. of 4 variables:
$ glyhb: Class 'labelled' atomic [1:379] 4.31 4.44 4.64 4.63 7.72 ...
 .. ..- attr(*, "label") = chr "Glycosolated Hemoglobin"

$ location: Factor w/ 2 levels "Buckingham","Louisa": 1 1 1 1 1 1 1 1 1 1 ...

$ gender: Factor w/ 2 levels "male","female": 2 2 2 1 1 1 1 1 1 2 ...

$ frame: Factor w/ 3 levels "small","medium": 2 3 3 3 2 3 2 2 3 1 ...

- attr(*, "na.action") = Class "omit" Named int [1:24] 44 51 60 64 65 70 109 110 111
117 ...
 .. ..- attr(*, "names") = chr [1:24] "44" "51" "60" "64" ...
```

Coding the Dependent Variable

In the absence of a true dichotomy for a dependent variable, for example, U.S. students versus international students, you will need to create one. I am interested in predicting diabetic condition given knowledge of *location, gender*, and *frame* (body size) in a logistic regression equation. I therefore must code *glyhb* into a dichotomous dependent variable (1 = *diabetic*, 0 = *not diabetic*) based on glycosolated hemoglobin levels >7, which would indicate *diabetic*, and values <7, indicating *not diabetic.* To recode *glyhb* into a dichotomous variable, we would use the *ifelse()* function to create two dichotomous values for a variable *depvar* and place them in the data set, *mydata.* The *ifelse()* function codes *depvar* = 1 if *glyhb* > 7, else *depvar* = 0. The R control structure command is as follows:

```
> mydata$depvar = ifelse(mydata$glyhb > 7, c("1"), c("0"))
The first few cases are printed again to show that the new variable
has been created and
takes on the intended values.
> head(mydata, n = 6)
  glyhb location  gender frame depvar
1 4.31 Buckingham female medium  0
2 4.44 Buckingham female large   0
3 4.64 Buckingham female large   0
4 4.63 Buckingham  male  large   0
5 7.72 Buckingham  male  medium  1
6 4.81 Buckingham  male  large   0
```

Checking for Nonzero Cell Counts

A final step prior to running our logistic regression equation is to check the data for nonzero cells in the crosstabs of variables. The *location, gender*, and *frame* predictor variables do not have 0 in their crosstab cells with the dependent variable *depvar.* The *xtabs()* function can be used to display the crosstab cell counts for the different variables.

```
> xtabs(~depvar + location, data = mydata)
location
depvar Buckingham Louisa
0 153 168
1 29 29

> xtabs(~depvar + gender, data = mydata)

gender
depvar male female
0      133    188
1       25     33

> xtabs(~depvar + frame, data = mydata)

frame
depvar small medium large
0       93     152     76
1        9      26     23
```

I went step by step through the data input, variable selection, missing data, recoding, and, finally, nonzero cell count processes to show the importance of screening and preparing data for statistical analysis. In the previous chapters, we also took steps and paid attention to some of these activities prior to conducting our statistical analysis. At some point, the phrase “Know your data” will ring true. If you prepare your data prior to statistical analysis and check the required assumptions, fewer problems will occur in the analysis, providing better sample estimates of the population parameters.

Logistic Regression Analysis

My logistic regression analysis can now proceed. First, I must declare categorical variables as *factors.* This requires the following R commands for each:

```
> mydata$location = factor(mydata$location)
> mydata$gender = factor(mydata$gender)
> mydata$frame = factor(mydata$frame)
> mydata$depvar = factor(mydata$depvar)
```

I am now ready to use the *glm()* function to run the logistic regression equation and the *summary()* function

to output the results.

```
> mylogit = glm (depvar ~ location + gender + frame, data = mydata,
family = "binomial")
> summary(mylogit)

Call:
glm(formula = depvar ~ location + gender + frame, family = "binomial",
 data = mydata)

Deviance Residuals:
   Min      1Q      Median     3Q     Max
-0.7489 -0.5728 -0.5448 -0.4363 2.2351

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.411885 0.440287   -5.478 0.000000043 ***
locationLouisa 0.006289 0.292557    0.021 0.98285
genderfemale   0.107978 0.300643    0.359 0.71948
framemedium    0.573023 0.408740    1.402 0.16094
framelarge     1.169743 0.432837    2.703 0.00688 **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 324.38 on 378 degrees of freedom
Residual deviance: 316.12 on 374 degrees of freedom
AIC: 326.12

Number of Fisher Scoring iterations: 5
```

The results indicated that *location* and *gender* differences were not statistically significant (p = .982 and p = .719, respectively). Regarding the *frame* variable, a large person was more indicative of diabetes than a medium or small person (p = .00688).

Confidence Intervals

A confidence interval can be computed using the standard error to form the confidence interval around these coefficients. The *confint.default()* function provides the p = .05 alpha-level results for a two-tailed percent (2.5% in each tail). The confidence intervals for *location* and *gender* contain the value of zero as expected since they were nonsignificant: Location CI = (-.567, +.579) and gender CI = (-.481, +.697). The medium-frame CI (-.228, +1.374) also contained the value of zero. The large-frame CI (.321, +2.018) did not contain the zero value, and as expected, it was statistically significant.

```
> confint.default(mylogit)

2.5 % 97.5 %
(Intercept) -3.2748312 -1.5489385
locationLouisa -0.5671126 0.5796900
genderfemale -0.4812705 0.6972275
framemedium -0.2280933 1.3741387
framelarge 0.3213982 2.0180873
```

Wald Chi-Square Test of Coefficient Significance

The Wald test in the *aod* package can be used to test the statistical significance of one or more (joint) coefficients, given their variance–covariance matrix. The arguments in the *wald.test()* function are listed as follows:

```
> wald.test(Sigma, b, Terms = NULL, L = NULL, H0 = NULL,
df = NULL, verbose = FALSE)
```

Sigma	A var-cov matrix, usually extracted from one of the fitting functions (e.g., lm, glm, ...).
b	A vector of coefficients with var-cov matrix Sigma. These coefficients are usually extracted from one of the fitting functions available in R (e.g., lm, glm,...).
Terms	An optional integer vector specifying which coefficients should be *jointly* tested, using a Wald *chi-squared* or *F* test. Its elements correspond to the columns or rows of the var-cov matrix given in Sigma. Default is NULL.

These can be printed out individually by the following commands:

```
> b = coef(mylogit)
> b
(Intercept) locationLouisa genderfemale framemedium framelarge
-2.411884853 0.006288706 0.107978490 0.573022739 1.169742731
```

and

```
> Sigma = vcov(mylogit)
> Sigma

(Intercept) locationLouisa genderfemale framemedium framelarge
(Intercept) 0.19385248 -0.050249138 -0.064462914 -0.123183051 -0.14672993
locationLouisa -0.05024914 0.085589632 0.004409604 -0.001355245 0.01401354
genderfemale -0.06446291 0.004409604 0.090386072 0.002975573 0.02493075
framemedium -0.12318305 -0.001355245 0.002975573 0.167068526 0.12248455
framelarge -0.14672993 0.014013535 0.024930754 0.122484551 0.18734769
```

To test the statistical significance of any given coefficient using the Wald chi-square test, we would select the row and column values for the *Terms* argument. To test the statistical significance of the *framelarge* variable level, it would be *Terms* = 5:5. This gives the same *p*-value result as in the *summary (mylogit)* function above.

```
> wald.test(b = coef(mylogit), Sigma = vcov(mylogit), Terms = 5:5)

Wald test:
----------

Chi-squared test:
X2 = 7.3, df = 1, P(> X2) = 0.0069
```

Our hand calculation verifies that the chi-square value is correct by using the *b* coefficient and variance term from the Sigma matrix in the formula:

$$X^2 = \frac{b^2}{(SE_b)^2} = \frac{(1.169742731)^2}{(.432837)^2} = \frac{1.368298}{0.18734769} = 7.30.$$

We can select different *Terms* for the other coefficients, thus obtaining the Wald chi-square test for each. The *p* values for each coefficient matches that in the *summary(mylogit)* function results above. These are computed below.

Wald Chi-Square Test for the Framemedium Coefficient

```
> wald.test(b = coef(mylogit), Sigma = vcov(mylogit), Terms = 4:4)

Wald test:
----------

Chi-squared test:
X2 = 2.0, df = 1, P(> X2) = 0.16
```

Wald Chi-Square Test for the Genderfemale Coefficient

```
> wald.test(b = coef(mylogit), Sigma = vcov(mylogit), Terms = 3:3)

Wald test:
----------

Chi-squared test:
X2 = 0.13, df = 1, P(> X2) = 0.72
```

Wald Chi-Square Test for the Location Coefficient

```
> wald.test(b = coef(mylogit), Sigma = vcov(mylogit), Terms = 2:2)

Wald test:
----------

Chi-squared test:
X2 = 0.00046, df = 1, P(> X2) = 0.98
```

Wald Chi-Square Test for the Intercept Coefficient

```
> wald.test(b = coef(mylogit), Sigma = vcov(mylogit), Terms = 1:1)
Chi-squared test:
X2 = 30.0, df = 1, P(> X2) = 0.000000043
```

The Wald chi-square tests are computed using the *b* coefficients divided by the square of their respective standard errors. The *coef()* and *vcov()* functions provide these values. The *Terms* argument specifies which squared standard error term is used for each regression coefficient. Personally, I would like to see these chi-square values in the table created by the *summary()* function.

Model Fit

Since *location* and *gender* were not statistically significant, we would drop them from the analysis and rerun the logistic regression equation. The variable *framelarge* was statistically significant (p = .00696), leaving a single predictor variable, *frame*, in the logistic regression equation. We will now continue with other results to provide additional information for interpreting the logistic regression model. The *summary()* function provides the following output:

```
> mylogit = glm (depvar ~ frame, data = mydata, family = "binomial")
> summary(mylogit)

Call:
glm(formula = depvar ~ frame, family = "binomial", data = mydata)

Deviance Residuals:
   Min     1Q     Median    3Q      Max
-0.7272 -0.5620 -0.5620 -0.4298 2.2035

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.3354    0.3491     -6.690  0.0000000000223 ***
framemedium  0.5696    0.4085      1.394  0.16325
framelarge   1.1401    0.4225      2.699  0.00696 **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 324.38 on 378 degrees of freedom
Residual deviance: 316.24 on 376 degrees of freedom

AIC: 322.24

Number of Fisher Scoring iterations: 5
```

Note: The residual deviance value is the −2LL value, which is 316.24.

Odds Ratio

Recall that the odds ratio effect size is computed as e^b, the exponent of each coefficient. This is computed by the R command

```
> exp(coef(mylogit))
(Intercept) framemedium framelarge
0.09677419 1.76754386 3.12719298
```

Confidence intervals can be placed around these odds ratio values as follows:

```
> exp(cbind(OR = coef(mylogit), confint(mylogit)))

Waiting for profiling to be done...

                    OR      2.5 %     97.5 %
(Intercept)  0.09677419 0.04537424  0.1812271
framemedium  1.76754386 0.82047057  4.1432752
framelarge   3.12719298 1.40744556  7.5042364
```

Many scholars find it hard to correctly interpret an odds ratio, but I continue the tradition of showing how to get the values. As noted earlier, the R *orsk* package will convert the odds ratio to the risk ratio for a probability interpretation. I turn my attention now to the important requirement of using and explaining the LL function.

Log Likelihood (-2LL), AIC and BIC

The LL value is obtained by the following *logLik()* function:

```
> logLik(mylogit)
'log Lik.' -158.1223 (df = 3)
```

You need to multiply the LL value by −2 to obtain the −2LL (residual deviance): −2 * −158.1223 = 316.245. Once we have a final logistic regression model, the −2LL value is used to compute the AIC index (AIC = −2 * LL + 2 * *df*). We can hand calculate this as follows: AIC = −2 * (-158.1223) + (2 * 3) = 322.2447. This is obtained using the R command

```
> AIC(mylogit)
[1] 322.2447
```

The BIC, which is related to the AIC, except for taking the log of the sample size, is computed as follows:

```
> n = nrow(mydata)
> BIC(mylogit)
[1] 334.0573
```

You can quickly see that the *AIC()* function can also obtain the BIC value by simply using the *k* argument to specify the log of the sample size:

```
> AIC(mylogit, k = log(n)) # same as BIC using log of sample size

[1] 334.0573
```

The large positive AIC value indicates that the overall model fit is statistically significant.

Chi-Square Difference Test

The chi-square difference test is yet another way to examine the overall model fit by comparing the null deviance and residual deviance values from the analysis (an intercept-only model vs. a model with one or more regression coefficients). The logistic regression results indicated a null deviance = 324.38 and a residual deviance = 316.24. The chi-square difference test shows the difference of 8.131173 as follows:

```
> with(mylogit, null.deviance - deviance)

[1] 8.131173
```

The degrees of freedom for the difference in the models is given by

```
> with(mylogit, df.null - df.residual)

[1] 2
```

The *p* value for the chi-square difference test is given by

```
> options(scipen = 999) # removes scientific notation

> with(mylogit, pchisq(null.deviance-deviance, df.null - df.residual, lower.tail = FALSE))

[1] 0.01715293
```

The chi-square difference test ($\Delta\chi^2 = 8.13$, $df = 2$, $p = .017$) is statistically significant. This indicates that *frame* is a statistically significant variable that adds variance explained above and beyond an intercept-only baseline (model). In practice, we start with the baseline model, then add variables, each time checking to see if the variable contributes above and beyond the intercept and/or previous variables in the model. This is why the order of the predictor variable entry in the logistic regression equation is important to consider and can lead to different results.

R-Squared Analog (Pseudovariance* R *Squared)

I am only reporting the calculations for the Cox and Snell *R*-squared and Nagelkerke *R*-squared values, because these two are the ones most often reported in other statistics packages. Recall that the formulas are as follows:

$$R^2_{CS} = 1 - \exp\left(\frac{-2LL_{model} - \left(-2LL_{baseline}\right)}{n}\right)$$

$$= 1 - \exp\left(\frac{324.376 - 316.245}{379}\right) = 1 - \exp\left(\frac{8.131}{379}\right) = .021$$

and

$$R_N^2 = \frac{R_{CS}^2}{1 - \exp\left(-\frac{-2LL_{Baseline}}{n}\right)} = \frac{.021}{1 - \exp\left(-\frac{316.245}{379}\right)}$$

$$= \frac{.021}{1 - \exp(-.8344)} = \frac{.021}{.5658} = .037.$$

The reported −2LL model is null deviance = 324.38 with 378 degrees of freedom (intercept-only model). The −2LL baseline is residual deviance = 316.24 with 376 degrees of freedom (intercept plus the *frame* predictor variable). Adding the one predictor variable lowered the −2LL value; the chi-square difference test above indicated a statistically significant change; and the Wald chi-square test indicated that *frame* was a statistically significant predictor variable. Finally, we have an effect size measure in the *R*-squared analog values. We can compute these two *R*-squared values as follows:

```
> Rcs = 1 - exp((mylogit$deviance - mylogit$null.deviance)/379)
> Rcs

[1] 0.02122578
```

and

```
> Rn = Rcs/(1-(exp(-(mylogit$null.deviance/379))))
> Rn

[1] 0.03690875
```

Note: The −2LLfull model (324.376) is the sum of 8.131 (chi-square for the one-predictor model) + 316.245 (-2LL intercept-only model). The *R*-squared values indicate a small effect size (.021-.037).

The power to control for a Type II error can be estimated using the *R*-squared analog values in the *pwr.f2.test()* function in the *pwr* package. We load the package and obtain the arguments as follows:

```
> library(pwr)
> ?pwr.f2.test

pwr.f2.test(u = NULL, v = NULL, f2 = NULL, sig.level = 0.05, power = NULL)
```

Arguments

u	degrees of freedom for numerator
v	degrees of freedomfor denominator
f2	effect size
sig.level	Significance level (Type I error probability)
Power	Power of test (1 minus Type II error probability)

The power for our example would have df_1 = 2 (p - 1 predictors) and df_2 = 379, which is specified and run using the Nagelkerke *R*-squared value as follows:

```
> pwr.f2.test(u = 2, v = 379, f2 = .0369, sig.level = .05, power = NULL)

Multiple regression power calculation

u = 2
v = 379
f2 = 0.0369
sig.level = 0.05
power = 0.9277476
```

So even though this is a small effect size (0.0369), we have sufficient power (.927) to detect the effect size. The determination of a small effect size is based on Cohen's rule of thumb (Cohen, 1988). It is quite possible that in the research literature, this effect size may be considered a moderate or even large effect size. It is best to determine the effect size found in your research area when possible.

Predicted Probabilities

The predicted probabilities are computed using the *predict()* function. I assigned the predicted probabilities to the variable *yhat.*

```
> yhat = predict(mylogit, newdata = mydata, type = "response")
```

Next, I add the predicted probabilities for *yhat* to the data set *mydata* and print out the first six rows of data using the *head()* function. This permitted me to see that the predicted values were included in the data set correctly.

```
> compdata = data.frame(mydata, yhat)
> head(compdata)
glyhb location gender frame depvar yhat
1 4.31 Buckingham female medium 1 0.1460674
2 4.44 Buckingham female large 1 0.2323232
3 4.64 Buckingham female large 1 0.2323232
4 4.63 Buckingham male large 1 0.2323232
5 7.72 Buckingham male medium 2 0.1460674
6 4.81 Buckingham male large 1 0.2323232
```

R-*Squared Computation*

It is easy to correlate the actual *Y* values with the predicted *Y* values using the *cor()* function once the predicted values are added to the data set. The *cor()* function will require the variable names from the *compdata* data set just created. You may recall that the variable names are referenced in a data set by the data set name, $, and the variable name, which is *compdata$depvar* for the actual *Y* values. We will need to convert the variable *depvar* back to a numeric variable (recall that we declared it as a *factor* variable before). The conversion uses the *as.numeric()* function. The R commands are as follows:

```
> compdata$depvar = as.numeric(compdata$depvar)
> r = cor(compdata$depvar, compdata$yhat)
> r
[1] 0.1468472
```

We now square the correlation to obtain the *R*-squared value for the logistic regression model.

```
> rsq = r^2
> rsq
[1] 0.02156411
```

This value resembles the Cox and Snell R_{CS}^2 value; it also indicates a small effect size.

Summary of Logistic Regression R Commands

I have put all of the R commands together in a logical progression to show the steps taken to conduct a logistic regression analysis. You can see that it requires diligence to make sure that all of the right steps are taken in the proper order. I have placed all of these commands in the R Logistic Regression script file (chap18.r), so that you should not have to manually enter them; copy them into your own R script file, add any additional R commands, and save for future use.

```
# Load Hmisc, get data set and attach data set
> library(Hmisc)
> getHdata(diabetes)
> attach(diabetes)

# Select variables from diabetes data set
> myvars = c("glyhb","location","gender","frame")
> newdata = diabetes[myvars]

# Omit missing values
> mydata = na.omit(newdata) # 379 obs 4 variables

# Code dichotomous dependent variable and add to data set
> mydata$depvar = ifelse(mydata$glyhb > 7, c("1"), c("0"))
> mydata # 379 obs 5 variables

# Check for non-zero cells
> xtabs(~depvar + location, data = mydata)
> xtabs(~depvar + gender, data = mydata)
> xtabs(~depvar + frame, data = mydata)

# Designate variables as factors
> mydata$location = factor(mydata$location)
> mydata$gender = factor(mydata$gender)
> mydata$frame = factor(mydata$frame)
> mydata$depvar = factor(mydata$depvar)

# Run logistic regression equation
> mylogit = glm (depvar ~ frame, data = mydata, family = "binomial")
> summary(mylogit)

# Compute R-squared analog values (Cox and Snell and Nagelkerke)
> Rcs = 1 - exp((mylogit$deviance - mylogit$null.deviance)/379)
> Rcs
> Rn = Rcs/(1-(exp(-(mylogit$null.deviance/379))))
> Rn
# Compute Predicted Probabilities and Add to Data Set
> yhat = predict(mylogit, newdata = mydata, type = "response")
> compdata = data.frame(mydata, yhat)
> head(compdata)

# Convert depvar to numeric, correlate depvar and yhat, square correlation
> compdata$depvar = as.numeric(compdata$depvar)
> r = cor(compdata$depvar, compdata$yhat)
>rsq = r^2
>rsq
```

Journal Article

Satcher and Schumacker (2009) published an article using binary logistic regression to predict modern homonegativity among 571 professional counselors. Counselors are ethically bound not to discriminate against persons based on sexual orientation (American Counseling Association. 2005). Sparse research, however, existed on the topic. Modern homonegativity is prejudices against gay men and lesbians based on current issues, such as equality and social justice, measured by the Modern Homonegativity Scale (Morrison & Morrison, 2002). Scores on the instrument were dichotomously coded into 0 = *low* and 1 = *high* modern homonegativity for the dependent variable based on the upper and lower quartiles—yielding 99 counselors with low modern homonegativity scores and 90 counselors with high modern homonegativity scores. Predictor variables that prior research had shown to be related to attitudes toward homosexuality were age, gender, race, political affiliation, education level, religious affiliation, personal contact or friendship, affirming information about homosexuality, and personal view regarding homosexuality having a biological origin.

The logistic regression equation predicted the probability of *Y* occurring for each person given responses on the independent predictor variables. In the logistic regression analysis, the observed and predicted values on *Y* were used to assess the fit of the equation using an LL statistic. The LL statistic indicated just how much unexplained variance was left after the regression model had been fitted—that is, large values indicated poor fitting models, while lower values indicated better fitting models. The first logistic regression model was a baseline model that only included the intercept value, which is based on the frequency of *Y* when *Y* = 1. The improvement in model fit was determined by a chi-square difference test, with 1 degree of freedom between successive models; this involved the subtraction of hypothesized new models with additional predictor variables from the baseline model:

$$\chi^2 = \text{LL(baseline model)} - \text{LL(Model A)}.$$

The individual contribution or statistical significance of independent predictor variables in the logistic regression equation was determined by computing a Wald statistic: Wald = B/SE_B, where B = regression coefficient and SE_B = standard error. The other approach used was to compare successive models, adding unique parameters each time and determining the reduction in −2LL and increase in pseudo *R*-squared and chi-square values, which indicated a better model fit and thus better classification and prediction of the probability of *Y.* Interpretation of the logistic regression coefficients, or e^B (exp(B)), indicated the change in odds, resulting in a unit change in the predictor variable. For a value greater than 1, the proportion of change in odds indicated that as the predictor variable values increased, the odds of the *Y* outcome increased. For a value less than 1, the proportion of change in odds indicated that as the predictor increased, the odds of the *Y* outcome decreased.

The baseline model, which only included the constant, yielded the −2LL statistic = 261.58 and percent classification = 52%. The final logistic regression equation, after adding successive predictor variables,

yielded the −2LL statistic = 32.97, percent classification = 85%, Nagelkerke *R*-squared = .659, and model χ^2 = 128.61.

The final binary logistic regression equation was as follows:

$$P(Y) = \frac{1}{1 + e^{-(-4.54 + 2.6(\text{attend}) + 1.77(\text{friend}) + 2.36(\text{political}) + 1.17(\text{trianing}) + 1.20(age))}}.$$

The probability of modern homonegativity was now predicted based on the final set of statistically significant predictor variables: attend church, friend, political affiliation, training, and age. For example, a counselor who attends church regularly (*attend* = 1), does not have a gay or lesbian friend (*friend* = 1), is a Republican (*political* = 1), has no training related to gay or lesbian sexual identity (*training* = 1), and is over the age of 48 years (*age* = 1) would have a 99% chance of having high modern homonegativity scores. In contrast, a counselor who does not attend church regularly (*attend* = 0), has a gay or lesbian friend (*friend* = 0), is a Democrat/other (*political* = 0), has training related to gay or lesbian sexual identity (*training* = 0), and is under the age of 48 years (*age* = 0) would have a 1% chance of having high homonegativity scores. The use of individual values that have different combinations of these predictor variables would lead to probability values between the 0 and 1 dependent variable values. The risk ratio was .99 divided by .01, which indicated that counselors are 99 times more likely to have high modern homonegativity scores with the predictor variable characteristics in the logistic regression model. Cohen (2000) clarifies the correct way to interpret odds ratio versus the probability ratio (risk ratio). Basically, an OR of 2:1 is not the same as having twice the probability of an event occurring. The use of the logistic regression equation to predict the probability of *Y* for any given individual based on the values for the predictor variables is recommended instead.

Conclusions

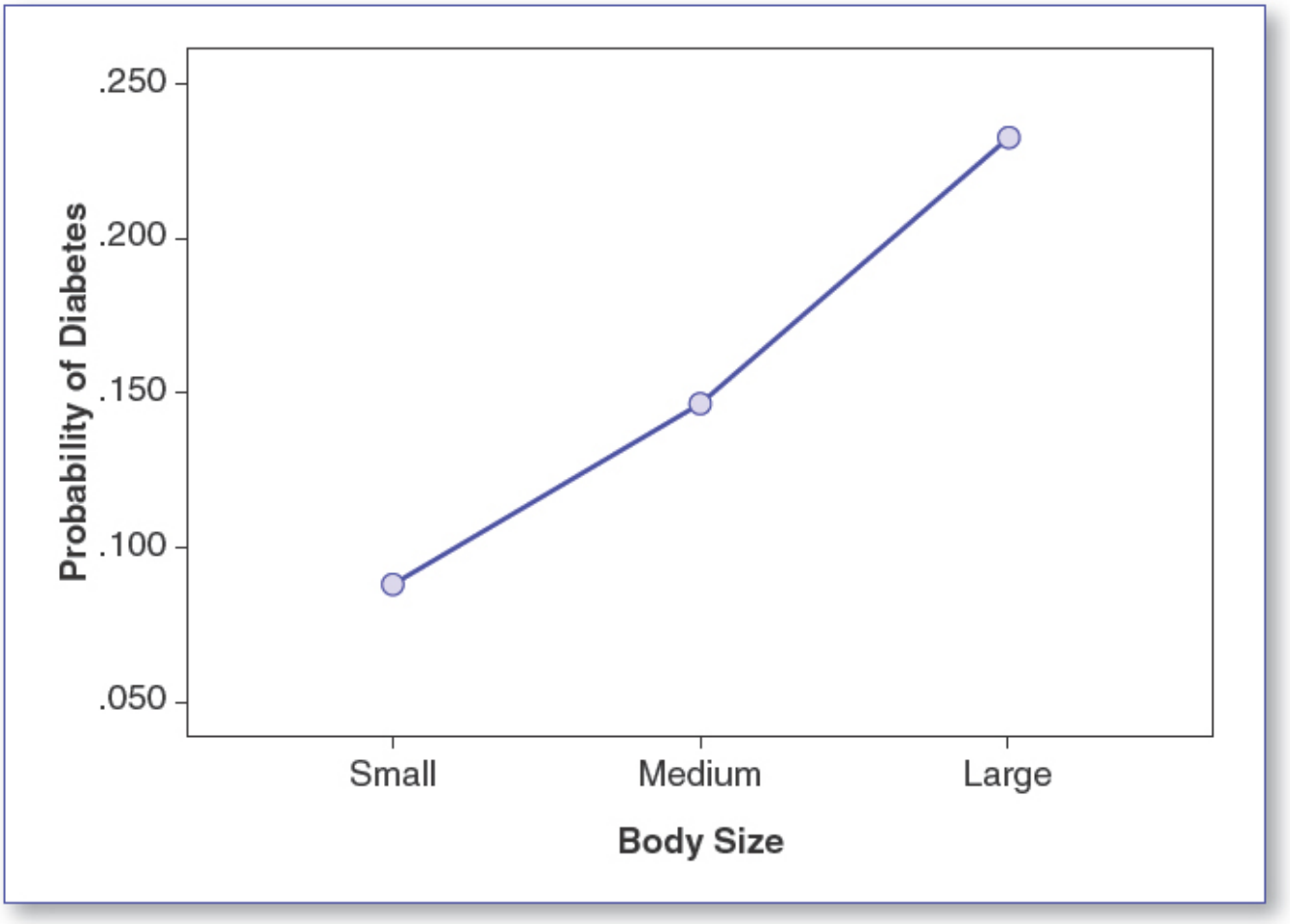
Logistic regression predicts a dichotomous dependent variable, rather than a continuous variable as in OLS regression. The assumptions in logistic regression do not require the dependent variable and the error variances (residuals) to be normally distributed and uncorrelated. There are other assumptions that must be met for logistic regression analyses, with the most severe violation occurring in the presence of nonzero cell counts. The logistic regression equation is also written in a multiplicative form due to exponentiation of the variable values. Finally, the overall model fit, significance of the regression weights, and effect size are computed using different values due to maximum likelihood estimation, rather than the least squares estimation used in OLS regression. For example, a chi-square test is used to indicate how well the logistic regression model fits the data.

The logistic regression analysis example indicated that *location* and *gender* were not statistically significant, but *frame* was at the $p < .05$ level of significance. An easy way to interpret a trend in the levels of this variable is to compute the number at each level who were diagnosed as being diabetic. If we compute the average on *Y* for those who were diagnosed as diabetic (*Y* = 1), the results will show that as a person's frame (body size)

increases, the probability of diabetes incidence too increases. When converting to an e^X value, a probability ratio interpretation is possible. Recall that the percent above a value of 1 indicates an increased probability. Having a large frame (body size) indicates a 26% greater chance of getting Type II diabetes, compared with a 15% chance with a medium frame and only a 9% chance with a small frame.

Frame	Total	Diabetic	Probability	e^X
Small	102	9	.088	1.09
Medium	178	26	.146	1.15
Large	99	23	.232	1.26

If we graph the probability values, a visual depiction shows this increase:



The logistic regression equation provides the computations to determine the significance of a predicator variable in terms of predicting the probability that $Y = 1$, which is referred to as $\hat{p}$. The probability that Y is 0 is $1 - \hat{p}$.

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = \alpha + \beta_1 X.$$

The *ln* symbol refers to a natural logarithm, and $\alpha + \beta_1 X$ is our familiar equation for the regression line. I showed how the predicted probabilities can be computed from the regression equation using the *predict()* function. So if we know the regression equation, we can calculate the probability that $Y = 1$ for a given value of X:

$$\hat{p} = \frac{\exp(\alpha + \beta_1 X)}{1 + \exp(\alpha + \beta_1 X)} = \frac{e^{\alpha + \beta_1 X}}{1 + e^{\alpha + \beta_1 X}}$$

The model fit and test of regression coefficient significance is best indicated by a chi-square difference test using −2LL values. We start with a baseline model (intercept-only), then add variables, each time determining if a better model fit occurs. This is similar to a change in the *R*-squared value explained in OLS regression modeling when another variable has been added to the equation. In logistic regression, we expect the −2LL values to decrease when adding variables that improve model fit. The chi-square difference test is not valid unless the two models compared involve one model that is a reduced form of the other model (nested).

Summary

This chapter covered a special type of multiple regression, binary logistic regression, where the dependent variable was measured at the dichotomous level rather than at the continuous level of measurement. I provided the set of special assumptions that affect the use of logistic regression, especially nonzero cell counts. The logistic regression equation and its special multiplicative form and maximum likelihood estimation were then presented. Model fit, tests for regression coefficient statistical significance, confidence interval, and effect size were discussed. These measures were different from the OLS multiple regression presented in the previous chapters. The logic of OLS multiple regression is similar, in that prediction of a dependent variable, determination of significant predictor variables, and model fit are desired, but the statistics used are different based on the binary outcome and maximum likelihood estimation. A binary logistic regression example was provided that took us step by step through the analyses using a set of data. A summary of the R commands used for the logistic regression analysis was provided to show the steps a researcher would take. I did not conduct the analyses in the five-step hypothesis testing approach, leaving that up to you to orchestrate.

The interpretation of logistic regression results from a published research article provided a meaningful end to the chapter. The application, prediction, and interpretation of the findings made it helpful to further understand when and how logistic regression analyses are used. It especially pointed out how different combinations of values for the independent predictor variables would lead to each individual predicted probability value. Beyond interpreting a person's individual probability of occurrence, logistic regression can provide a measure of group membership or classification accuracy. The statistical significance between the actual and predicted group membership can be tested using the chi-square statistic.

This chapter did not cover the many other types of multiple regression equations, for example, Poisson, multinominal, ordinal, probit, longitudinal growth, or mixed models. The treatment of these multiple regression models is beyond the scope of this book. I do, however, cover another special type of multiple regression in the next chapter—that is, log-linear regression models based on frequency counts for the dependent variable.

Exercises

1. List four important assumptions in logistic regression.
    a.
    b.
    c.
    d.
2. Given a logistic regression equation with a dependent variable *saver* (0 = *does not save money on a regular basis*, 1 = *saves money on a consistent basis*) and three predictor variables, age (0 = *under* 30, 1 = *over 30*), education (0 = *high school*, 1 = *college*), and income (0 = *less than $50,000*, 1 = *more than $50,000*), the following logistic regression coefficients were reported:

Variable	Coefficient
Intercept	-1.00
Age	1.50
Education	1.15
Income	2.50

    a. State the form of the logistic regression equation using these coefficients.
    b. Predict the probability value for an individual with the following characteristics: over 30, college educated, with income more than $50,000.
    c. Predict the probability value for an individual with the following characteristics: over 30, high school education, with income less than $50,000.
    d. Comparing these two individuals, which one is more likely to be a saver?
    e. Which variable is the determining factor that affects this percent difference the most?
3. Suppose a researcher came to you for help in interpreting her or his logistic regression output. The statistician had run a baseline model with −2LL = 240 and *df* = 4 but then showed another regression model with two predictor variables having −2LL = 340 and *df* = 2. The researcher wanted you to tell whether the two models were statistically significant.
    a. What would you compute?
    b. What are the reported values?

c. What would you conclude given the results?

True or False Questions

T	F	a. All independent predictor variables in a logistic regression must be continuous.
T	F	b. A logistic regression equation predicts individual values between 0 and 1.
T	F	c. The constant term, α, in the logistic regression equation defines the baseline model.
T	F	d. The regression coefficient, b, in the logistic regression equation implies the log odds corresponding to a one-unit change in the variable X_i, controlling for other variables in the equation.
T	F	e. The chi-square difference test should not be used because it leads to erroneous conclusions.
T	F	f. The *R*-squared analog tests indicate the statistical significance of the regression coefficients.
T	F	g. The logistic regression model involving the predictor variable *X* binary coded (0,1) is given by $P(Y) = \alpha + \beta X$.

Web Resources

Chapter R script file is available at http://www.sagepub.com/schumacker

Logistic Regression R script file: chap18.r
orsk package: http://CRAN.R-project.org/package=orsk

http://dx.doi.org/10.4135/9781506300160.n18