Multinomial Logistic Regression

Dr. Jon Starkweather and Dr. Amanda Kay Moske

Multinomial logistic regression is used to predict categorical placement in or the probability of category membership on a dependent variable based on multiple independent variables. The independent variables can be either dichotomous (i.e., binary) or continuous (i.e., interval or ratio in scale). Multinomial logistic regression is a simple extension of binary logistic regression that allows for more than two categories of the dependent or outcome variable. Like binary logistic regression, multinomial logistic regression uses maximum likelihood estimation to evaluate the probability of categorical membership.

Multinomial logistic regression does necessitate careful consideration of the sample size and examination for outlying cases. Like other data analysis procedures, initial data analysis should be thorough and include careful univariate, bivariate, and multivariate assessment. Specifically, multicollinearity should be evaluated with simple correlations among the independent variables. Also, multivariate diagnostics (i.e. standard multiple regression) can be used to assess for multivariate outliers and for the exclusion of outliers or influential cases. Sample size guidelines for multinomial logistic regression indicate a minimum of 10 cases per independent variable (Schwab, 2002).

Multinomial logistic regression is often considered an attractive analysis because; it does not assume normality, linearity, or homoscedasticity. A more powerful alternative to multinomial logistic regression is discriminant function analysis which requires these assumptions are met. Indeed, multinomial logistic regression is used more frequently than discriminant function analysis because the analysis does not have such assumptions. Multinomial logistic regression does have assumptions, such as the assumption of independence among the dependent variable choices. This assumption states that the choice of or membership in one category is not related to the choice or membership of another category (i.e., the dependent variable). The assumption of independence can be tested with the Hausman-McFadden test. Furthermore, multinomial logistic regression also assumes non-perfect separation. If the groups of the outcome variable are perfectly separated by the predictor(s), then unrealistic coefficients will be estimated and effect sizes will be greatly exaggerated.

There are different parameter estimation techniques based on the inferential goals of multinomial logistic regression analysis. One might think of these as *ways of applying* multinomial logistic regression when strata or clusters are apparent in the data.

Unconditional logistic regression (Breslow & Day, 1980) refers to the modeling of strata with the use of dummy variables (to express the strata) in a traditional logistic model. Here, one model is applied to all the cases and the stata are included in the model in the form of separate dummy variables, each reflecting the membership of cases to a particular stata.

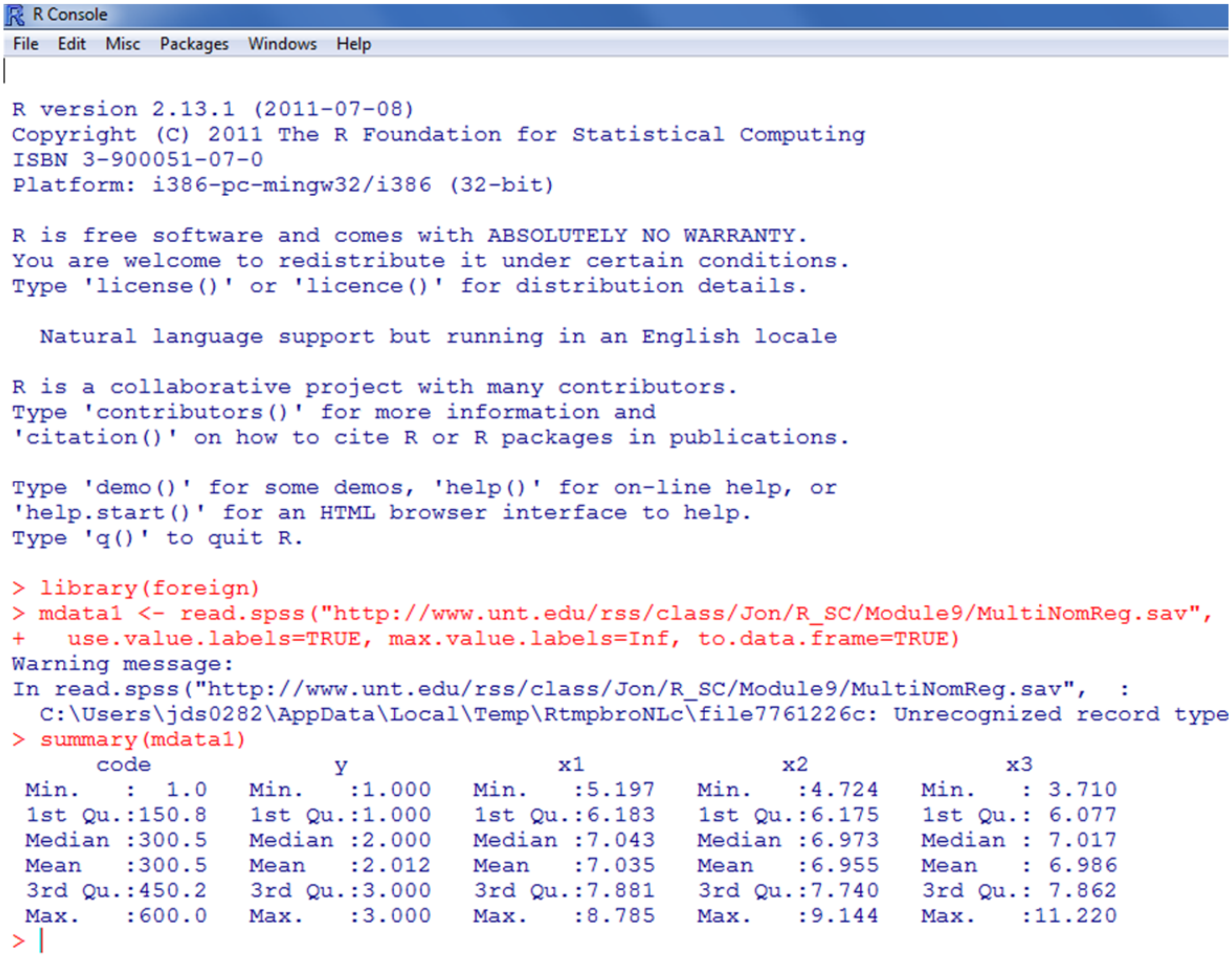
Conditional logistic regression (Breslow & Day, 1980; Vittinghoff, Shiboski, Glidden, & McCulloch, 2005) refers to applying the logistic model to each of the stata individually. The coefficients of the predictors (of the logistic model) are *conditionally* modeled based on the membership of cases to a particular stata.

Marginal logistic modeling (Vittinghoff, Shiboski, Glidden, & McCulloch, 2005) refers to an aggregation of the stata so that the coefficients reflect the population values averaged across the stata. As a rudimentary example, consider averaging each of the conditional logistic coefficients, from the previous paragraph, to arrive at set marginal coefficients for all members of the population – regardless of stata membership.

Variable selection or model specification methods for multinomial logistic regression are similar to those used with standard multiple regression; for example, sequential or nested logistic regression analysis. These methods are used when one dependent variable is used as criteria for placement or choice on subsequent dependent variables (i.e., a decision or flow-chart). For example, many studies indicate the decision to use drugs follows a sequential pattern, with alcohol at an initial stage followed by the use of marijuana, cocaine, and other illicit drugs.

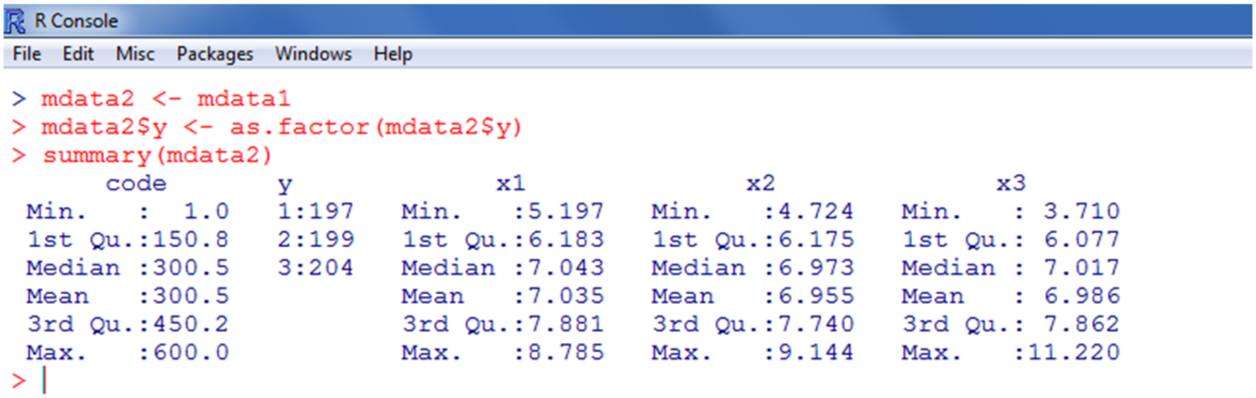
Example

For the following example a fictitious data set will be used. The data includes a single categorical dependent variable with three categories. The data also includes three continuous predictors. The data contained enough cases (*N* = 600) to satisfy the cases to variables assumption mentioned earlier. First, import the data using the ‘foreign’ package and get a summary.



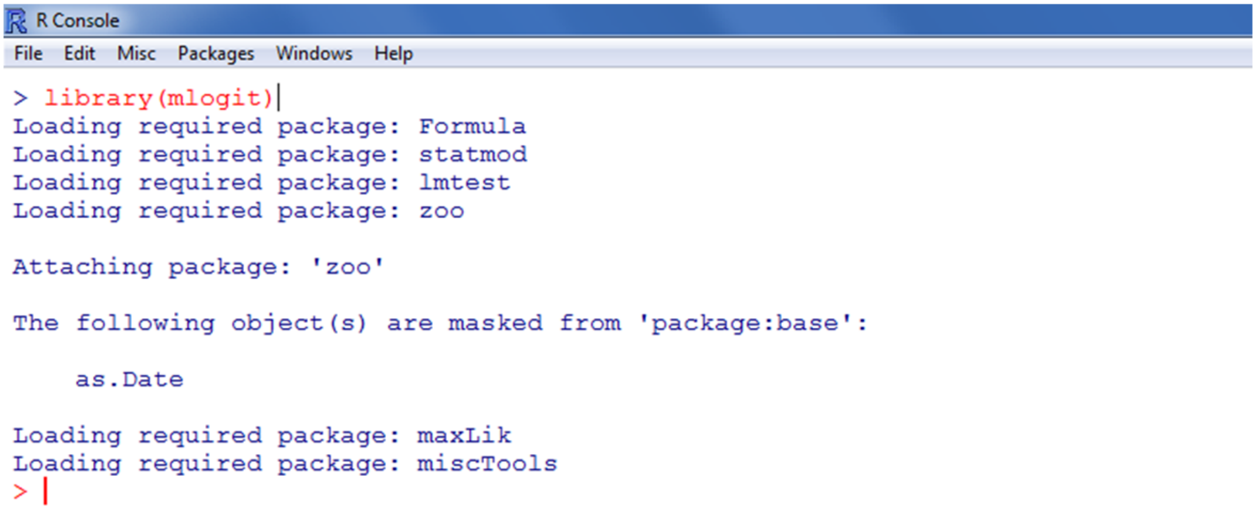
```
R version 2.13.1 (2011-07-08)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-pc-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> library(foreign)
> mdata1 <- read.spss("http://www.unt.edu/rss/class/Jon/R_SC/Module9/MultiNomReg.sav",
+   use.value.labels=TRUE, max.value.labels=Inf, to.data.frame=TRUE)
Warning message:
In read.spss("http://www.unt.edu/rss/class/Jon/R_SC/Module9/MultiNomReg.sav",  :
  C:\Users\jds0282\AppData\Local\Temp\RtmpbroNLc\file7761226c: Unrecognized record type
> summary(mdata1)
      code             y               x1              x2              x3        
 Min.   :  1.0   Min.   :1.000   Min.   :5.197   Min.   :4.724   Min.   : 3.710  
 1st Qu.:150.8   1st Qu.:1.000   1st Qu.:6.183   1st Qu.:6.175   1st Qu.: 6.077  
 Median :300.5   Median :2.000   Median :7.043   Median :6.973   Median : 7.017  
 Mean   :300.5   Mean   :2.012   Mean   :7.035   Mean   :6.955   Mean   : 6.986  
 3rd Qu.:450.2   3rd Qu.:3.000   3rd Qu.:7.881   3rd Qu.:7.740   3rd Qu.: 7.862  
 Max.   :600.0   Max.   :3.000   Max.   :8.785   Max.   :9.144   Max.   :11.220  
> 
```

Next, we need to identify the outcome variable as a factor (i.e. categorical).



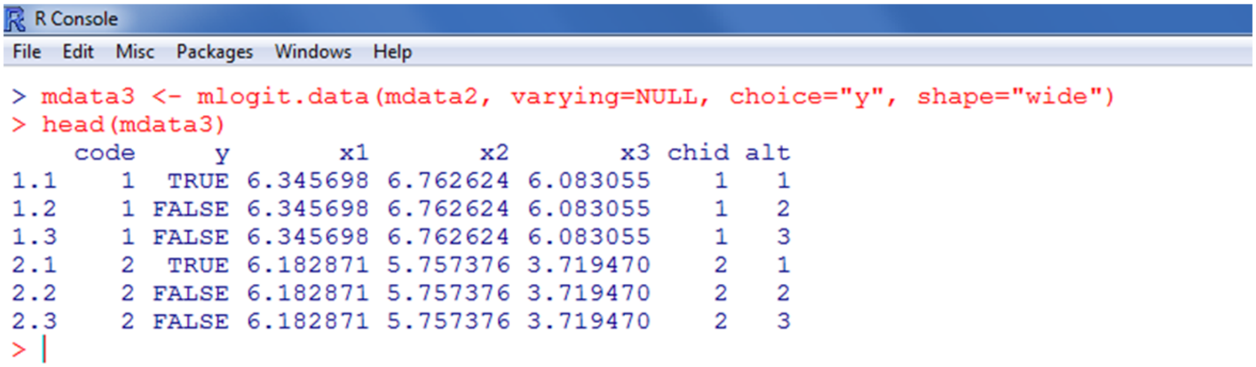
```
> mdata2 <- mdata1
> mdata2$y <- as.factor(mdata2$y)
> summary(mdata2)
      code          y            x1              x2              x3
 Min.   :  1.0   1:197   Min.   :5.197   Min.   :4.724   Min.   : 3.710
 1st Qu.:150.8   2:199   1st Qu.:6.183   1st Qu.:6.175   1st Qu.: 6.077
 Median :300.5   3:204   Median :7.043   Median :6.973   Median : 7.017
 Mean   :300.5           Mean   :7.035   Mean   :6.955   Mean   : 6.986
 3rd Qu.:450.2           3rd Qu.:7.881   3rd Qu.:7.740   3rd Qu.: 7.862
 Max.   :600.0           Max.   :8.785   Max.   :9.144   Max.   :11.220
>
```

Next, we need to load the 'mglogit' package (Croissant, 2011) which contains the functions for conducting the multinomial logistic regression. Note, the 'mlogit' packages requires six other packages.



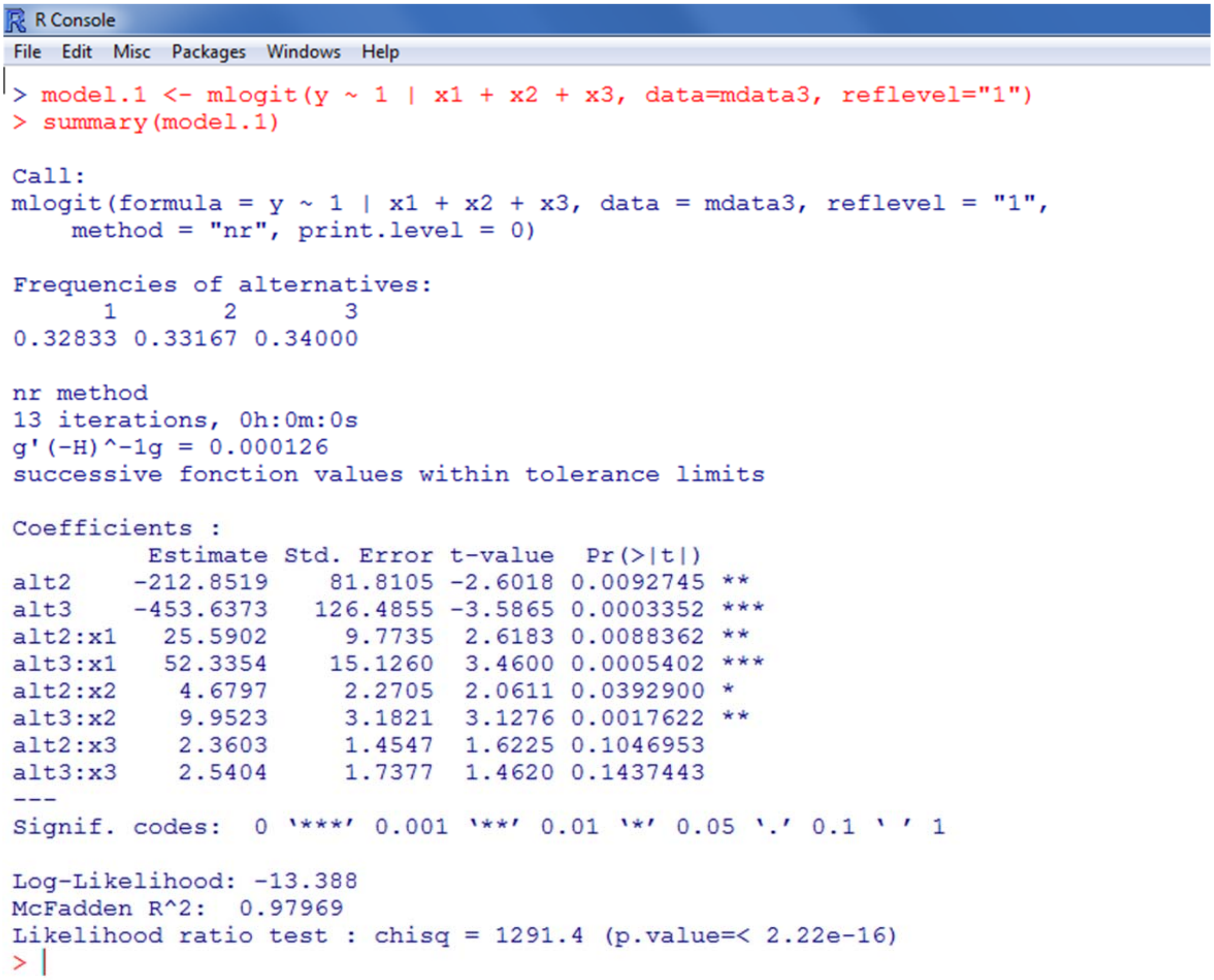
```
> library(mlogit)
Loading required package: Formula
Loading required package: statmod
Loading required package: lmtest
Loading required package: zoo

Attaching package: 'zoo'

The following object(s) are masked from 'package:base':

    as.Date

Loading required package: maxLik
Loading required package: miscTools
>
```

Next, we need to modify the data so that the multinomial logistic regression function can process it. To do this, we need to expand the outcome variable (y) much like we would for dummy coding a categorical variable for inclusion in standard multiple regression.



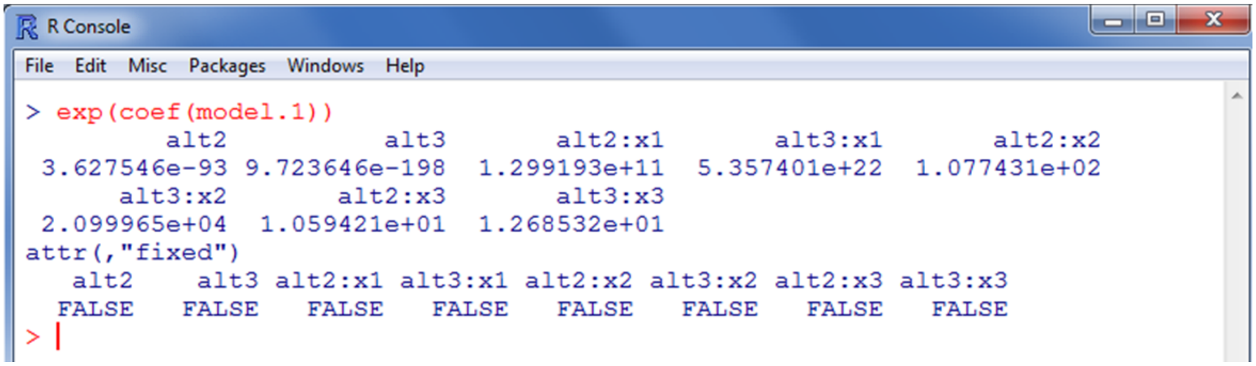
```
> mdata3 <- mlogit.data(mdata2, varying=NULL, choice="y", shape="wide")
> head(mdata3)
    code     y       x1       x2       x3 chid alt
1.1    1  TRUE 6.345698 6.762624 6.083055    1   1
1.2    1 FALSE 6.345698 6.762624 6.083055    1   2
1.3    1 FALSE 6.345698 6.762624 6.083055    1   3
2.1    2  TRUE 6.182871 5.757376 3.719470    2   1
2.2    2 FALSE 6.182871 5.757376 3.719470    2   2
2.3    2 FALSE 6.182871 5.757376 3.719470    2   3
>
```

Now we can proceed with the multinomial logistic regression analysis using the ‘mlogit’ function and the ubiquitous ‘summary’ function of the results. Note that the reference category is specified as “1”.



```
> model.1 <- mlogit(y ~ 1 | x1 + x2 + x3, data=mdata3, reflevel="1")
> summary(model.1)

Call:
mlogit(formula = y ~ 1 | x1 + x2 + x3, data = mdata3, reflevel = "1",
    method = "nr", print.level = 0)

Frequencies of alternatives:
      1       2       3
0.32833 0.33167 0.34000

nr method
13 iterations, 0h:0m:0s
g'(-H)^-1g = 0.000126
successive fonction values within tolerance limits

Coefficients :
          Estimate Std. Error t-value  Pr(>|t|)
alt2     -212.8519    81.8105 -2.6018 0.0092745 **
alt3     -453.6373   126.4855 -3.5865 0.0003352 ***
alt2:x1    25.5902     9.7735  2.6183 0.0088362 **
alt3:x1    52.3354    15.1260  3.4600 0.0005402 ***
alt2:x2     4.6797     2.2705  2.0611 0.0392900 *
alt3:x2     9.9523     3.1821  3.1276 0.0017622 **
alt2:x3     2.3603     1.4547  1.6225 0.1046953
alt3:x3     2.5404     1.7377  1.4620 0.1437443
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Log-Likelihood: -13.388
McFadden R^2:  0.97969
Likelihood ratio test : chisq = 1291.4 (p.value=< 2.22e-16)
> 
```

The results show the logistic coefficient (B) for each predictor variable for each alternative category of the outcome variable; alternative category meaning, not the reference category. The logistic coefficient is the expected amount of change in the logit for each one unit change in the predictor. The logit is what is being predicted; it is the odds of membership in the category of the outcome variable which has been specified (here the first value: 1 was specified, rather than the alternative values 2 or 3). The closer a logistic coefficient is to zero, the less influence the predictor has in predicting the logit. The table also displays the standard error, *t* staistic, and the p-value. The *t* test for each coefficient is used to determine if the coefficient is significantly different from zero. The Pseudo R-Square (McFadden R^2) is treated as a measure of effect size, similar to how R^2 is treated in standard multiple regression. However, these types of metrics do not represent the amount of variance in the outcome variable accounted for by the predictor variables. Higher values indicate better fit, but they should be interpreted with caution. The Likelihood Ratio chi-square test is alternative test of goodness-of-fit. As with most chi-square based tests however, it is prone to inflation as sample size increases. Here, we see model fit is significant $\chi^2 = 1291.40$, $p < .001$, which indicates our full model predicts significantly better, or more accurately, than the null model. To be clear, you want the p-value to be less than your established cutoff (generally 0.05) to indicate good fit. To get the expected B values, we can use the ‘exp’ function applied to the coefficients.



```
> exp(coef(model.1))
         alt2          alt3        alt2:x1        alt3:x1        alt2:x2 
 3.627546e-93 9.723646e-198   1.299193e+11   5.357401e+22   1.077431e+02 
      alt3:x2        alt2:x3        alt3:x3 
 2.099965e+04   1.059421e+01   1.268532e+01 
attr(,"fixed")
   alt2    alt3 alt2:x1 alt3:x1 alt2:x2 alt3:x2 alt2:x3 alt3:x3 
  FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE   FALSE 
> 
```

The Exp(B) is the odds ratio associated with each predictor. We expect predictors which increase the logit to display Exp(B) greater than 1.0, those predictors which do not have an effect on the logit will display an Exp(B) of 1.0 and predictors which decease the logit will have Exp(B) values less than 1.0. Keep in mind, the first two listed (alt2, alt3) are for the intercepts.

Further reading on multinomial logistic regression is limited. Several authors (Garson, 2006; Mertler & Vannatta, 2002; Pedhazur, 1997) provide discussions of binary logistic regression in the context of graduate level textbooks, which provides insight into multinomial because it is a direct extension. Clearly those authors believe that if one is inclined to understand binary logistic, then one is also likely to understand multinomial logistic. There is merit in this position because one is an extension of the other and both use maximum likelihood (an ogive function). However; other authors provide either direct examples of multinomial logistic regression (Schwab, 2002; Tabachnick & Fidell, 2001) or a full discussion of multinomial logistic regression (Aldrich & Nelson, 1984; Fox, 1984; Hosmer & Lemeshow, 1989; Menard, 1995).

Until next time, *you can tell everybody this is your song…*

References & Resources

Aldrich, J. H., & Nelson, F. D. (1984). Linear probability, logit, and probit models. Thousand Oaks, CA: Sage.

Breslow, N. E., & Day, N. E. (1980). Statistical Methods in Cancer Research. Lyon, UK: International Agency for Research on Cancer.

Croissant, Y. (2011). Package ‘mlogit’. http://cran.r-project.org/web/packages/mlogit/index.html

Fox, J. (1984). Linear statistical models and related methods: With applications to social research. New York: Wiley.

Garson, G. D. (2011). “Logistic Regression”, from Statnotes: Topics in Multivariate Analysis. http://faculty.chass.ncsu.edu/garson/pa765/statnote.htm.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). Multivariate Data Analysis (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Hoffmann, J. (2003). Generalized linear models: An applied approach. Boston, MA: Allyn & Bacon.

Hosmer, D. W., & Lemeshow, S. (1989). Applied logistic regression. New York: Wiley.

Menard, S. (1995). Applied logistic regression analysis. Thousand Oaks, CA: Sage.

Mertler, C. & Vannatta, R. (2002). Advanced and multivariate statistical methods (2nd ed.). Los Angeles, CA: Pyrczak Publishing.

Pedhazur, E. J. (1997). Multiple regression in behavioral research: Explanation and prediction (3rd ed.). New York: Harcourt Brace.

Schwab, J. A. (2002). Multinomial logistic regression: Basic relationships and complete problems. http://www.utexas.edu/courses/schwab/sw388r7/SolvingProblems/

Tabachnick, B. G. & Fidell, L. S. (2001). Using multivariate statistics (4th ed.). Needleham Heights, MA: Allyn and Bacon.

Vittinghoff, E., Shiboski, S. C., Glidden, D. V., & McCulloch, C. E. (2005). Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models. New York: Springer Science+Business Media, Inc.