Class Six: Multinomial and ordinal logiistic regression

Xiaoyan Wang

Feb 22, 2018

Learning objectives

- Understand what multinomial and ordinal logitstic regressions are
- Know the assumptions of multinomial and ordinal logitstic regressions
- Be able to use R to perform multinomial and ordinal logistic regression
- Know how to perform hypothesis testing
- Know how to estimate and interpret odds ratio
- Know how to predict and interpret the marginal effects

What is multinomial logistic regression (MLR)?

- A form of linear regression analysis conducted when the dependent variable is nominal with more than two levels (e.g., insurance status, occupation)
- Used to describe data and to explain the relationship between one dependent nominal variable and one or more independent variables.

Crucial features

- Estimates a separate binary logistic regression model for each dummy variables, therefore the result is J-1 binary logistic regression models, suppose our dependent nominal variable has J levels
- Each model conveys the effect of predictors on the probability of success in that category, in comparison to the reference category.

Link function

- Take the logarithm of the odds of the probability of y=m versus the probability of y=b, where b is the base outcome whose coefficients associated with the independent variables are constrained to be zero.

Assumptions

- Dependent variable is nominal with more than two levels (can also be ordinal variables)
- One or more independent variables that are continuous, ordinal or nominal
- Independence of the dependent variable (Hausman-McFadden test)
- Independence of observations
- No multicollinearity
- A linear relationship between any *continuous* independent variables and the logit transformation of the dependent variable.
- No outliers or highly influential points (Pregibon's (1981) deltabeta or Cook's distance)
- Unlike logistic regression where there are many statistics for performing model diagnostics, it is not as straightforward to do diagnostics with multinomial logistic regression models.

MLR vs. binary logistic regression

- Binary logit (2 nominal outcomes): we have 2-1 or one β vector (i.e., parameters showing the effect of each predictor on the logit), including one intercept term $\beta0$
- Multinomial logit (J nominal outcomes): we have J-1 sets of parameters, or J-1 sets of βm (for m=1, 2, … J-1). Each βm contains a non-zero intercept $\beta0$
- You can view binary logit as a special case of multinomial logit model in which J=2

Example

- Breast cancer data downlaoded from Surveillance, Epidemiology, and End Results Program (SEER), a premier data source for cancer statistics in the US. SEER collects incidence, prevalence and survival data.
- Patients with primary breast cancer, aged between 18-65 years and diagnosed with Stage I to IV were included.
- In this study, we are interested in the effects of insurance status on the stage at diagnosis among breast cancer patients.
- The covariates are comprised of patients' demographics.

Get the data and load libraries

```
rm(list=ls())
#install.packages("nnet")
#install.packages("MASS")
#install.packages("erer")
#install.packages("readr")
#install.packages("knitr")
#install.packages("tidyverse")
library(readr) #for read txt file
library(knitr) #for creating nicer tables
library(tidyverse) # for various packages
```

```
## -- Attaching packages ----------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 2.2.1     v purrr   0.2.4
## v tibble  1.4.2     v dplyr   0.7.4
## v tidyr   0.8.0     v stringr 1.2.0
## v ggplot2 2.2.1     v forcats 0.2.0
```

Data Management

```
##Look at the data
names(Breast_SEER_Class6)<-c("Age", "Race", "Sex", "Diagnosis_year", "Stage", "First", "Pati
table(Breast_SEER_Class6$Insur)
```

```
## 
##             Any Medicaid Insurance status unknown                   Insured 
##                     5342                      922                     25983 
##     Insured/No specifics                Uninsured 
##                     3336                      783 
```

```
table(Breast_SEER_Class6$Stage)
```

```
## 
##          0        IA        IB       IIA       IIB      IIIA      IIIB 
##         32     15928       827      7828      4515      2459       702 
##       IIIC    IIINOS        IV UNK Stage 
##        866        29      2033      1121 
```

```
table(Breast_SEER_Class6$Sex)
```

Get the complete data: drop the cases who were did not have first primary diagnosis of breast cancer and with missing values on any variables used

```
Data1 <- Data %>%
filter(First_cat!="No"&!is.na(Stage_cat)&!is.na(Insur_cat)&!is.na(Black_cat)&!is.na(Age_num)
```

Execute a mutilnomial logistic regression

Below we use the multinom function from the *nnet* package to estimate a multinomial logistic regression model.

```
##Re-leveling data, choose stage IV as reference
attach(Data1)
Stage_cat_re <- relevel(Stage_cat, ref = "StageIV")
##Execute a mutilnomial regression with insurance as independent variable and demographics a
mod <- multinom(Stage_cat_re ~ Insur_cat + Age_num + Male_cat + Black_cat)
```

```
## # weights:  28 (18 variable)
## initial  value 41353.160792
## iter  10 value 34593.988823
## iter  20 value 32759.082902
## final  value 32632.669527
## converged
```

```
summary(mod)
```

Interpretations

- *Continous variable*
- A one-year increase in age is associated with the increase in the log odds of being in stage I vs. stage IV in the amount of .021, holding other variables constant.
- *Categorical variable*
- The log odds of being in stage I vs. stage IV will increase by 1.492 if changing from no insurance to private insurance, holding other variables constant.
- To know the effect of *Insured* on the logit of a non-base outcome (e.g.,stage II) relative to another non-base outcome (e.g.,stage III), you take the difference of the two *coefficients* (0.8579903-0.6583665=0.1996238).
- Meaning: the log odds of being in stage II vs. stage III will increase by 0.200 if changing from no insurance to private insurance, holding other variables constant.

Hypothesis test

- Use Z-test to test individual parameter. – The p-value from such test indicates whether or not we can reject the null hypothesis of the parameter=0 relative to the effect on the base outcome at a given level of statistical significance.

```
z <- summary(mod)$coefficients/summary(mod)$standard.errors
# 2-tailed Z test
p <- (1 - pnorm(abs(z), 0, 1)) * 2
p
```

```
##           (Intercept) Insur_catMedicaid Insur_catInsured      Age_num
## StageI   6.547635e-01        0.01048902     0.000000e+00 1.800871e-11
## StageII  0.000000e+00        0.17030744     5.324852e-11 5.153443e-04
## StageIII 5.581029e-07        0.25162579     3.313087e-05 3.969484e-09
##          Male_catMale Black_catBlack
## StageI   4.305358e-06   0.000000e+00
## StageII  6.609725e-02   9.197487e-11
## StageIII 1.810542e-01   3.747187e-05
```

Hypothesis test: Wald test and likelihood ratio test

- Test whether an independent variable is significant for the entire model with all outcomes (not from a single comparison).
- Test a group of parameters (i.e., multiple independent variables).
- Test whether or not two outcomes can be combined, that is to test the null hypothesis that the two categories are indistinguishable.

Odds ratio

```
## extract the coefficients from the model and exponentiate
exp(coef(mod))
```

```
##          (Intercept) Insur_catMedicaid Insur_catInsured   Age_num
## StageI     0.9119598          1.445865         4.444783 1.0208695
## StageII    6.0927526          1.206380         2.358416 0.9894019
## StageIII   3.2206251          1.209652         1.931634 0.9795025
##          Male_catMale Black_catBlack
## StageI      0.2894788      0.4986219
## StageII     0.6266689      0.6504227
## StageIII    0.6589386      0.7230239
```

```
#Example
str(summary(mod))
```

```
## List of 31
##  $ n              : num [1:3] 6 0 4
##  $ nunits         : int 11
##  $ nconn          : num [1:12] 0 0 0 0 0 0 0 0 7 14 ...
```

Interpretations

- *Continous variable*
- The odds of having stage I breast cancer relative to stage IV are 1.02 times greater with one-year increases in age, holding other variables constant.
- *Categorical variable*
- The odds of having stage I breast cancer relative to stage IV are 4.44 times greater for privately insured patients than uninsured cases, holding other variables constant.

Important notes about interpretation

- “Using odds ratios to interpret the logit model is very common, but rarely is it sufficient for understanding the results of the model. We strongly prefer methods of interpretation that are based on predicted probabilities.” (Long & Freese, 2014, p.227)
- Marginal effects show the change in probability when the predictor or independent variable increases by one unit, holding all other independent variables constant at specific values. For continuous variables this represents the instantaneous change given that the ‘unit’ may be very small. For binary variables, the change is from 0 to 1, so one ‘unit’ as it is usually thought.

Three approaches showing marginal effects

- Marginal effect at the mean (MEM): Compute the marginal effect of one independent variable with all other independent variables held at their means.
- Marginal effect at representative values (MER): Compute the marginal effect of x with variables held at specific values that are selected for being especially instructive for the substantive questions being considered. The MEM is a special case of the MER.
- Average marginal effect (AME): Compute the marginal effect of independent variable for each observation at its observed values, and then compute the average of these effects.
- No single approach can meet all needs. The best should be a combination of all three that help addresses the research questions.

Marginal effect at the mean (MEM)

- Due to the variable types inconsistence (e.g.,the mean value of Male variable may be not meaningful), however we can calculate here by hand.
- The structural model
- Suppose J=5, the model can be expressed as:

Marginal effect at representative values (MER)

```
MER<- data.frame(Insur_cat = "Insured", Age_num = 50, Male_cat="Male",Black_cat="Black")
predict(mod, newdata = MER, "probs")
```

```
##   StageIV    StageI   StageII  StageIII
## 0.1401811 0.2303644 0.4819420 0.1475125
```

Overview of ordinal outcome variables

- Key features of an ordinal variable: it does have a ranking order, but distances between levels are not measureable
- Alternative models
- Binary logistic regression – collapse categories to create a dichotomous outcome
- Multinomial logistic regression – ignore the ranking order by treating the outcome as a nominal variable

Ordinal logistic regression

- Ordinal regression is used to predict the dependent variable with 'ordered' multiple categories and independent variables.

Assumptions

- Dependent variable should be ordinal variables with more than two levels.
- One or more independent variables that are continuous, ordinal or categorical
- There is no multicollinearity.
- Have proportional odds or parallel regression.

Execute a mutilnomial logistic regression

```
mod1 <- polr(Stage_cat ~ Insur_cat + Age_num + Male_cat + Black_cat, Hess=TRUE)
summary(mod1)
```

```
## Call:
## polr(formula = Stage_cat ~ Insur_cat + Age_num + Male_cat + Black_cat,
##     Hess = TRUE)
##
## Coefficients:
##                      Value Std. Error t value
## Insur_catMedicaid -0.19755   0.079987  -2.470
## Insur_catInsured  -0.79789   0.076110 -10.483
## Age_num           -0.02755   0.001299 -21.205
## Male_catMale       0.76243   0.141069   5.405
## Black_catBlack     0.35840   0.032644  10.979
##
## Intercepts:
##                  Value    Std. Error t value
## StageI|StageII    -2.1790   0.0997   -21.8543
## StageII|StageIII  -0.3523   0.0988    -3.5642
## StageIII|StageIV   0.7988   0.1003     7.9640
```

Hypothesis test

- Use Z-test to test individual parameter
- The p-value from such test indicates whether or not we can reject the null hypothesis of parameter=0 at a given level of statistical significance.

```
## store table
ctable <- coef(summary(mod1))
## calculate and store p values
p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2
p
```

```
## Insur_catMedicaid  Insur_catInsured          Age_num      Male_catMale
##      1.351736e-02      1.028754e-25     8.512785e-100      6.493769e-08
##     Black_catBlack    StageI|StageII  StageII|StageIII  StageIII|StageIV
##      4.831746e-28     7.075016e-106      3.650220e-04      1.665806e-15
```

```
## combined table
ctable <- cbind(ctable, "p value" = p)
```

Odds ratio and 95%CIs

```
## odds ratios
exp(cbind("Odds ratio" = coef(mod1), confint.default(mod1, level = 0.95)))
```

```
##                    Odds ratio     2.5 %    97.5 %
## Insur_catMedicaid  0.8207359 0.7016471 0.9600372
## Insur_catInsured   0.4502766 0.3878777 0.5227137
## Age_num            0.9728283 0.9703545 0.9753084
## Male_catMale       2.1434786 1.6257007 2.8261661
## Black_catBlack     1.4310322 1.3423400 1.5255846
```

Interpretation

- Exp(B) above was estimated, meaning compare odds from low to high.
- Examples
- *Continous variable*
- The odds of having stage I versus combined outcomes stage II to IV are 0.973 times greater with one-year increase in age, holding other variables constant.
- *OR*
- The odds of having stage I and II versus combined outcomes stage III and IV are 0.973 times greater with one-year increase in age, holding other variables constant.
- *Categorical variable*
- The odds of having stage I versus combined outcomes stage II to IV are 0.450 times greater for patients with private insurance than the ones without insurance, holding other variables constant.

Marginal effect at the mean (MEM)- calculate by hand based on the structural function

- Suppose dependent variable has 3 levels, the model can be expressed as:

Marginal effect at representative values (MER)

```
MER<- data.frame(Insur_cat = "Insured", Age_num = 50, Male_cat="Male",Black_cat="Black")
predict(mod1, newdata = MER, "probs")
```

```
##    StageI   StageII  StageIII   StageIV
## 0.2451597 0.4235151 0.1958349 0.1354903
```

Average marginal effect (AME)

```
AME <- ocME(mod1)
AME
```

```
##                    effect.StageI effect.StageII effect.StageIII
## Insur_catMedicaid          0.049         -0.025          -0.015
## Insur_catInsured           0.191         -0.068          -0.071
## Age_num                    0.007         -0.003          -0.002
## Male_catMale              -0.178          0.053           0.071
## Black_catBlack            -0.088          0.038           0.030
##                    effect.StageIV
## Insur_catMedicaid          -0.009
## Insur_catInsured           -0.052
## Age_num                    -0.001
## Male_catMale                0.054
## Black_catBlack              0.020
```

Interpretation

- Using the average marginal effects and holding all other variables at a constant level, we find that:
- The probability of having stage IV for insured patients is lowerer than that for uninsured patients by 0.052;
- Every one-year increase in patients' age increases the probability of having stage I by 0.007.