Eco311 Optional Reading: Multinomial Logistic Regression (MLR)

(Jing Li, Miami University)

1. We can use Logistic Regression when the outcome or dependent variable takes only two categories. Examples are Employed vs Unemployed, and Trump vs Biden. Multinomial Logistic Regression (MLR) is needed if there are more than two categories.

2. For instance, in the data we will use, the outcome variable y is insure, which takes three categories of Indemnity, Prepaid, and Uninsure.

```
> library(readxl)
> setwd("/Users/lij14/Dropbox")
> data = read_excel("mlogitdata.xls")
> table(is.na(data$insure))
FALSE  TRUE
  616    28
> table(data$insure)
Indemnity   Prepaid  Uninsure
      294       277        45
```

   There are 28 missing values for insure; among the 616 non-missing values, 297 are Indemnity, 277 are Prepaid, and 45 are Uninsure. We wonder whether the predictor nonwhite matters for insure.

3. The **multinom** function in the **nnet** package can be used to estimate the MLR:

```
> install.packages("nnet")
> library(nnet)
> model = multinom(insure ~ nonwhite, data = data)
> summary(model)
Coefficients:
         (Intercept)  nonwhite
Prepaid   -0.1879116 0.6608144
Uninsure  -1.9419427 0.3780860

Std. Errors:
```

```
          (Intercept)  nonwhite
Prepaid    0.09376432 0.2157328
Uninsure   0.17821926 0.4075742

Residual Deviance: 1103.567
AIC: 1111.567
```

4. Just like a logistic regression, MLR is fitted by maximum likelihood method. The distribution for the i-th observation is

$$P(y_i = Prepaid) = \frac{exp(\beta_0^{Prepaid} + \beta_1^{Prepaid} nonwhite)}{1 + exp(\beta_0^{Prepaid} + \beta_1^{Prepaid} nonwhite) + exp(\beta_0^{Uninsure} + \beta_1^{Uninsure} nonwhite)} \quad (1)$$

$$P(y_i = Uninsure) = \frac{exp(\beta_0^{Uninsure} + \beta_1^{Uninsure} nonwhite)}{1 + exp(\beta_0^{Prepaid} + \beta_1^{Prepaid} nonwhite) + exp(\beta_0^{Uninsure} + \beta_1^{Uninsure} nonwhite)} \quad (2)$$

$$P(y_i = Indemnity) = \frac{1}{1 + exp(\beta_0^{Prepaid} + \beta_1^{Prepaid} nonwhite) + exp(\beta_0^{Uninsure} + \beta_1^{Uninsure} nonwhite)} \quad (3)$$

We can verify that each probability is bounded between 0 and 1, and their sum is equal to one.

5. Notice that there are two intercepts $\beta_0^{Prepaid}, \beta_0^{Uninsure}$, and two slopes $\beta_1^{Prepaid}, \beta_1^{Uninsure}$. The interpretation is based on the log odds:

$$log\left(\frac{P(y_i = Prepaid)}{P(y_i = Indemnity)}\right) = \beta_0^{Prepaid} + \beta_1^{Prepaid} nonwhite \quad (4)$$

$$log\left(\frac{P(y_i = Uninsure)}{P(y_i = Indemnity)}\right) = \beta_0^{Uninsure} + \beta_1^{Uninsure} nonwhite \quad (5)$$

So the log odds of Prepaid relative to Indemnity is $\beta_0^{Prepaid} = -0.1879116$ when nonwhite is zero. When nonwhite changes from 0 to 1, the log odds of Prepaid relative to Indemnity rises by $\beta_1^{Prepaid} = 0.6608144$. Moreover, the log odds of Uninsure relative to Indemnity is $\beta_0^{Prepaid} = -1.9419427$ when nonwhite is zero. When nonwhite changes from 0 to 1, the log odds of Uninsure relative to Indemnity rises by $\beta_1^{Uninsure} = 0.3780860$.

6. To sum up, for a white person (nonwhite is zero), the two negative intercepts imply that $P(y_i = Prepaid) < P(y_i = Indemnity)$ and $P(y_i = Uninsure) < P(y_i = Indemnity)$. So a white person is more likely to choose Indemnity. For a black person (nonwhite is

one), the two positive slopes imply that the probability of choosing Prepaid or Uninsure relative to Indemnity rises.

7. We can verify this finding by **table** function

```
> table(data$insure[data$nonwhite==0])
Indemnity   Prepaid  Uninsure
      251       208        36

> table(data$insure[data$nonwhite==0])/length(data$insure[data$nonwhite==0])
 Indemnity    Prepaid   Uninsure
0.48455598 0.40154440 0.06949807

> table(data$insure[data$nonwhite==1])
Indemnity   Prepaid  Uninsure
       43        69         9

> table(data$insure[data$nonwhite==1])/length(data$insure[data$nonwhite==1])
 Indemnity    Prepaid   Uninsure
0.34126984 0.54761905 0.07142857

> log(0.40154440/0.48455598)
[1] -0.1879149
> log(0.06949807/0.48455598)
[1] -1.941934
```

We see the change in probability of choosing Prepaid across race (from 0.40154440 to 0.54761905) is substantial; while the change in probability of choosing Uninsure is marginal (from 0.06949807 to 0.07142857). That explains the t value for $\beta_1^{Prepaid} = 0.6608144/0.2157328 > 1.96$ is significant, but the t value for $\beta_1^{Uninsure} = 0.3780860/0.4075742 < 1.96$ is not. The log odds are the same as the intercepts reported before.

8. We get the same results by maximizing a user-defined log likelihood function

```
> # user-defined log likelihood
> data = data[!is.na(data$insure),]
```

```
> cat("sample size is", nrow(data), "\n")
sample size is 616
> data$y1 = (data$insure=="Prepaid")
> data$y2 = (data$insure=="Uninsure")
> data$y3 = 1-data$y1-data$y2
>
> fmullogliklogit = function(b) {
+ zz1 = b[1]+data$nonwhite*b[2]
+ zz2 = b[3]+data$nonwhite*b[4]
+ p1 = exp(zz1)/(1+exp(zz1)+exp(zz2))
+ p2 = exp(zz2)/(1+exp(zz1)+exp(zz2))
+ p3 = 1/(1+exp(zz1)+exp(zz2))
+ return(-sum(data$y1*log(p1)+data$y2*log(p2)+data$y3*log(p3)))
+ }
> optim(c(1,0,1,0), fmullogliklogit,method="BFGS")
$par
[1] -0.1879186  0.6607970 -1.9419690  0.3783258
$value
[1] 551.7835
```

9. We can also get the same results by running two logistic regressions: one compares Prepaid to Indemnity; the other compares Uninsure to Indemnity:

```
> # alternatively, run two logistic regressions
> datas1 = data[data$y2==0,]
> coef(glm(formula = y1~nonwhite, family = "binomial",data=datas1))
(Intercept)    nonwhite
 -0.1879149   0.6608212
> datas2 = data[data$y1==0,]
> coef(glm(formula = y2~nonwhite, family = "binomial",data=datas2))
(Intercept)    nonwhite
 -1.9419340   0.3779585
```

10. Note that we exclude Uninsure when running the first logistic regression. This is called Independence of Irrelevant Alternatives (IIA) assumption. Google to learn more.