Multinomial Logistic Regression
R-library needed: VGAM
data file: brand_preference.csv

- In this case, the qualitative DV, instead of being a binary variable, can take K values.
- Example: DV = brand preference
  IV = gender, age, education

Sample data = $\{(X_{1i}, X_{2i}, \ldots, X_{Pi}, Y_i), i = 1, 2, \ldots, n\}$

In the case of binary logistic regression, the qualitative DV (Y) is either 0 or 1 , i.e., $Y_i = 1$ if i-th response is Success, and 0 otherwise.

In situations where the qualitative DV can take K values, we can express Y_i in terms of K binary responses $Y_{i1}, Y_{i2}, \ldots, Y_{iK}$ where

$$Y_{ij} = \begin{cases} 1 \text{ if i-th response is categoty j} \\ 0 \qquad\qquad\qquad \text{otherwise} \end{cases}$$

Since the response for case i can only be in one category j,

$$\sum_{j=1}^{K} Y_{ij} = 1 \text{ for each i.}$$

Recall that, for binary logistic regression, we model the logit

$$\ln\left(\frac{p}{1-p}\right) = \ln\left(\frac{p_1}{1-p_1}\right) = \ln\left(\frac{p_1}{p_2}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_P X_P = X^T \beta_{12}$$

where

$p_1 = P(Y = 1),\ p_2 = P(Y = 2).$

For K categories, there are K(K-1)/2 pairs, and therefore K(K-1)/2 linear models would have to be fit. This is avoided by arbitrarily making one of the categories (say K-th) the baseline and fitting K-1 logit models.

$$\ln\left(\frac{p_1}{p_K}\right) = \beta_{10} + \beta_{11} X_1 + \ldots + \beta_{1P} X_P$$

$$\ln\left(\frac{p_2}{p_K}\right) = \beta_{20} + \beta_{21} X_1 + \ldots + \beta_{2P} X_P$$

...

$$\ln\left(\frac{p_{K-1}}{p_K}\right) = \beta_{K0} + \beta_{K1} X_1 + \ldots + \beta_{KP} X_P$$

It can be shown that

$$p_j = \frac{\exp^{\beta_{j0} + \beta_{j1} X_1 + \ldots + \beta_{jP} X_P}}{1 + \sum_{k=1}^{K-1} e^{\beta_{k0} + \beta_{k1} X_1 + \ldots + \beta_{kP} X_P}}$$

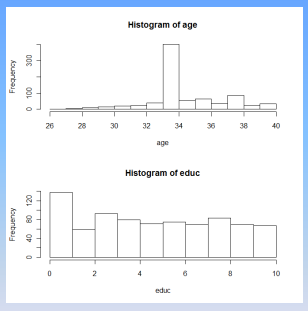
NOTE:
(1) As in the case of logistic regression, the β's are estimated by the method of maximum likelihood.
(2) For each X, there will be K β's.

Example:
DV = preference for hotel brand
IV = gender, age, education (years of education past high school)

```
xx <- read.csv("K:/DataMining/Data/brand_preference.csv",
header=TRUE)
names(xx)
[1] "female" "age"    "educ"   "hotel"
```


Descriptive Statistics of data

```
female
  0   1
298 506
```




```
table(hotel)
hotel
  1   2   3
232 340 232
```

```
# install package VGAM
library(VGAM)
m.out <- vglm(hotel~female+age+educ, family=multinomial(), na.action=na.pass)
summary(m.out)
```

Call:
vglm(formula = hotel ~ female + age + educ, family = multinomial(), na.action = na.pass)

Pearson Residuals:

	Min	1Q	Median	3Q	Max
log(mu[,1]/mu[,3])	-5.6858	-0.44984	-0.32336	0.68053	7.7123
log(mu[,2]/mu[,3])	-4.8460	-0.67938	-0.44118	0.95906	1.8578

Coefficients:

	Value	Std. Error	t value
(Intercept):	1 24.2960242	2.172153	11.18523
(Intercept):	2 11.7712627	1.556724	7.56156
female:1	-0.4648455	0.214926	-2.16281
female:2	0.0855761	0.189146	0.45243
age:1	-0.6923699	0.062235	-11.12517
age:2	-0.3244444	0.043282	-7.49610
educ:1	0.0097041	0.033808	0.28704*
educ:2	0.0103743	0.029177	0.35557*

* Educ is not significant (indicated by small magnitudes of t-values) – run model without educ.

Number of linear predictors: 2

Names of linear predictors: log(mu[,1]/mu[,3]), log(mu[,2]/mu[,3])

Dispersion Parameter for multinomial family: 1

Residual Deviance: 1544.739 on 1600 degrees of freedom

Log-likelihood: -772.3694 on 1600 degrees of freedom

Number of Iterations: 5


```
m.out2 <- vglm(hotel~female+age, family=multinomial(), na.action=na.pass)
summary(m.out2)
```

Call:
vglm(formula = hotel ~ female + age, family = multinomial(), na.action = na.pass)

Pearson Residuals:

	Min	1Q	Median	3Q	Max
log(mu[,1]/mu[,3])	-5.7433	-0.44984	-0.32533	0.67856	7.7985
log(mu[,2]/mu[,3])	-4.9030	-0.68165	-0.44449	0.94920	1.8132

Coefficients:

	Value	Std. Error	t value
(Intercept):1	24.354096	2.164362	11.25232
(Intercept):2	11.832644	1.547953	7.64406
female:1	-0.464434	0.214904	-2.16112
female:2	0.086084	0.189120	0.45518
age:1	-0.692652	0.062236	-11.12941
age:2	-0.324726	0.043277	-7.50348

$$\log\left(\frac{P(hotel=1)}{P(hotel=3)}\right) = 24.35 - .46\,female - .69age$$

$$\log\left(\frac{P(hotel=2)}{P(hotel=3)}\right) = 11.83 + .09\,female - .32age$$

Number of linear predictors: 2

Names of linear predictors: log(mu[,1]/mu[,3]), log(mu[,2]/mu[,3])

Dispersion Parameter for multinomial family: 1

Residual Deviance: 1544.874 on 1602 degrees of freedom

Log-likelihood: -772.437 on 1602 degrees of freedom

Number of Iterations: 5

Interpretation of multinomial logistic regression coefficients

With 1 unit change in Age, log of P(Y=1)/P(Y=3) decreases by 0.69
log of P(Y=2)/P(Y=3) decreases by .32

Exponentiate the coefficients:

```
exp(coef(m.out2))
(Intercept):1     (Intercept):2     female:1          female:2
 3.774414e+10  1.376740e+05  6.284910e-01  1.089898e+00

age:1                  age:2
 5.002475e-01          7.227252e-01
```

With 1 unit change in Age, P(Y=1)/P(Y=3) decreases by 0.50
P(Y=2)/P(Y=3) decreases by 0.72

For a binary predictor such as female:

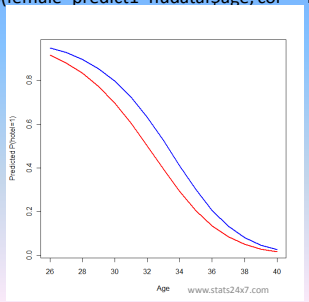
P(Y=1)/P(Y=3) for females (compared to males) is 0.63
P(Y=2)/P(Y=3) for females (compared to males) is 1.09

Graphing multinomial logistic regression results

```
nudatam <- data.frame(female=0, age=seq(26,40, 1))
male_predict <- predict(m.out2, newdata = nudatam, type="response")
nudataf <- data.frame(female=1, age=seq(26,40, 1))
female_predict <- predict(m.out2, newdata = nudataf, type="response")
```

	male_predict		
	1	2	3
1	0.94893382	0.04940055	0.001665625
2	0.92702440	0.06972287	0.003252726
3	0.89631858	0.09739456	0.006286859
4	0.85396579	0.13406055	0.011973658
5	0.79691970	0.18074375	0.022336544
6	0.72270019	0.23680731	0.040492501
7	0.63075561	0.29859757	0.070646816
8	0.52415605	0.35848746	0.117356488
9	0.41056429	0.40567921	0.183756498
10	0.30100145	0.42969310	0.269305453
11	0.20614582	0.42516013	0.368694046
12	0.13236428	0.39440010	0.473235618
13	0.08030981	0.34571881	0.573971380
14	0.04649828	0.28918749	0.664314227
15	0.02594381	0.23311200	0.740944187


```
male_predict1 <- male_predict[,1]
plot(male_predict1~nudatam$age, type="l",col="blue",lwd=2,ylab
= "Predicted P(hotel=1)", xlab="Age")
female_predict1 <- female_predict[,1]
lines(female_predict1~nudataf$age, col = "red", lwd=2)
```



REMARKS ABOUT MULTINOMIAL LOGISTIC REGRESSION

- Diagnostics are complicated.
- Requires large sample sizes.
- Small # of observations in a cell may cause model to not run, or give very poor results. One can do a crosstab between Y and categorical predictors to find out if such is the case.