

Ordinal Logistic Regression

- OLR is used when Y is ordinal – for example, Y is a measure of customer ratings of a restaurant on a 5-point Likert scale.

We will need to install and load the R-library 'Design'. The library Hmisc needs to be installed as well for 'Design' to work.

Let

Y = customer response (ordinal)

Y_C = continuous customer response (latent variable)

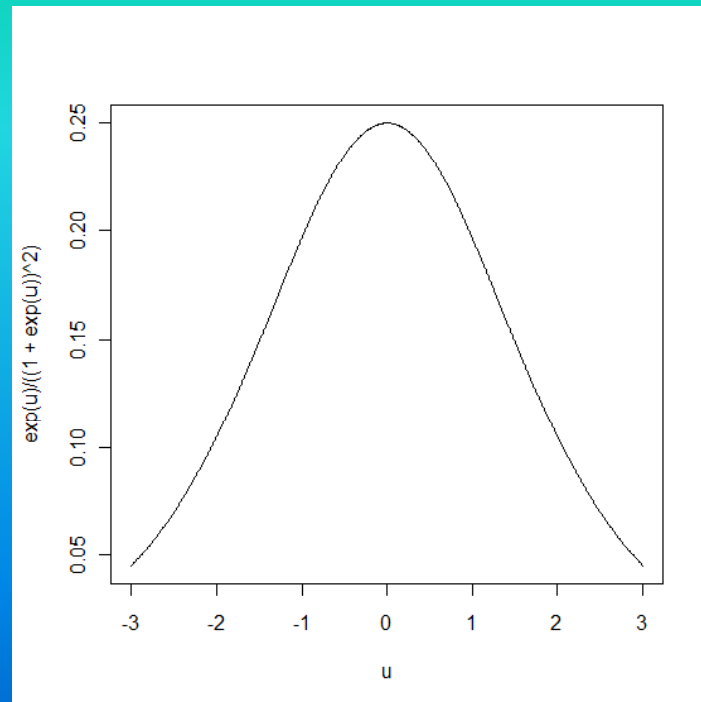
X_1 = predictor (assume just one predictor for now)

$$Y_{Ci} = \beta^*_0 + \beta^*_1 X_{1i} + k e$$

error $e \sim$ logistic distribution (see graph on next slide)

with mean 0, sd σ_C

$$\begin{aligned} P(Y \leq 1) &= P(Y_C \leq a_1) = P(\beta^*_0 + \beta^*_1 X_{1i} + k e \leq a_1) \\ &= P(e \leq (a_1 - \beta^*_0)/k - (\beta^*_1/k)) \\ &= P(e \leq \alpha_1 + \beta_1 X_1) \\ &= \exp(\alpha_1 + \beta_1 X_1) / [1 + \exp(\alpha_1 + \beta_1 X_1)] \end{aligned}$$



Similarly,

$$P(Y \leq 2) = \frac{\exp(\alpha_2 + \beta_1 X_1)}{1 + \exp(\alpha_2 + \beta_1 X_1)}$$

NOTE: Intercepts are different, but slopes are same. This is called *the proportional odds* model.

As before, ML method is used to estimate the parameters.

Example: A fast food chain wants to determine what factors affect size of soda (small, medium or large) customers order: factors being considered are

X_1 = type of food ordered (chicken sandwich or hamburger)

X_2 = fries ordered or not

X_3 = age of customer

The data file `soda_pref.csv` has the data collected on Y and the three predictors.

```
xx <- read.csv("K:/DataMining/Data/soda_pref.csv", header=TRUE)
attach(xx)
```

```
names(xx)
[1] "pref" "food" "fries" "age"
```

Descriptive Statistics

```
table(pref)
pref
 1    2    3
239 150  44

table(fries)
fries
 0  1
373 60
```

```
table(food)
food
 0    1
361  72
```

```
summary(age)
  Min. 1st Qu. Median Mean 3rd Qu. Max.
 20.00 21.00 21.00 21.03 21.00 22.00
```

```
xtabs(~food+pref)
  pref
food 1 2 3
 0 215 117 29
 1 24 33 15
```

```
xtabs(~fries+pref)
  pref
fries 1 2 3
 0 206 134 33
 1 33 16 11
```

Many researchers use OLS regression: this is not recommended since OLS assumptions are violated when DV is not an interval scale variable.

Ordinal Logistic Regression should be used – provided some of the cells do not have very low frequency counts.

Xtabs on previous slide show this is not the case.

```
library(Design)
```

```
# the following two lines of code needed for Design  
predictors <- datadist(food, fries, age)  
options(datadist = 'predictors')
```

```
ologit1 <- lrm(pref~food+fries+age, data = xx, na.action = na.pass)
```

```
# na.action = na.pass tells R to skip missing data
```

```
> ologit1
```

Logistic Regression Model

```
lrm(formula = pref ~ food + fries + age, data = xx, na.action =  
na.pass)
```

Frequencies of Responses

```
 1  2  3  
239 150 44
```

Obs	Max Deriv	Model L.R.	d.f.	P	C	Dxy
433	1e-10	27.13	3	0	0.607	0.214
	Gamma	Tau-a	R2	Brier		
	0.322	0.121	0.072	0.234		

Coef	S.E.	Wald	Z	P
y>=2	-12.116328	4.2666	-2.84	0.0045
y>=3	-14.185910	4.2810	-3.31	0.0009
food	1.033740	0.2496	4.14	0.0000
fries	-0.004363	0.2872	-0.02	0.9879
age	0.558108	0.2030	2.75	0.0060

χ^2 - Likelihood Ratio value of 27.13 is significant (i.e., fitted model is better than no-predictor model)

fries is not significant, food and age are.

1 unit increase in age increases log-odd ratio of Y by .56.

STATS24x7.com© 2010 ADI-NV, INC

11

exp(ologit1\$coefficients)

y>=2	y>=3	
5.469473e-06	6.904583e-07	
food	fries	age
2.811563e+00	9.956469e-01	1.747364e+00

For 1 unit increase in age, the odds of the low and middle categories of pref versus the high category are 1.74 times larger (all other variables held constant).

Proportional odds assumption implies that same is true (1.75 times increase) for low pref and combined middle+high pref.

STATS24x7.com© 2010 ADI-NV, INC

12