Statistics with R – Advanced Level

Section 2

Predictive Techniques

Lesson 11 - Binomial regression

```
mobi <- read.csv("mobilenet.csv")

View(mobi)

#########
### how to perform the binomial logistic regression
#########

#########
### Basic assumptions:

# the independent variables do not present outliers
# there is no important multicollinearity
##########

### we will predict the chance that a subject uses mobile
Internet
### based on the other three variables

### dependent variable: has/does not have mobile internet
(1/0)
### explainers: income, hours spent on the Internet, where
they use the Internet
```

```
### where is a dichotomous variable: at home (0) and at the
office (1)

### very important: the dependent variable must be coded
numerically

### to run the regression, we use the glm function

model <- glm(mobile~income+hours+where, data=mobi,
family=binomial())

### for the categorical independent variable where,
### home (0) is the reference category

summary(model)

### Null deviance - the difference between the LL of the
saturated model
### and the LL of the null model (with intercept only)

### Residual deviance - difference between the LL of the
### saturated model and the LL of the proposed model

### the saturated model is the model where each case has
its own parameter
### the proposed model is better than the saturated model
### because it has a lower LL

######### compute the antilogarithms of the coefficients
######### these antilogs actually represent the chance that
a subject uses mobile Internet

expb <- exp(coef(model))

print(expb)

### compute the confidence interval of the antilogarithms

intexp <- exp(confint(model))

print(intexp)
```

Lesson 12 - Binomial regression - goodness-of-fit measures

```
mobi = read.csv("mobilenet.csv")

View(mobi)

#########
### the binomial logistic regression - goodness-of-fit
indicators
#########

### run the regression model again

model <- glm(mobile~income+hours+where, data=mobi,
family=binomial())

### compute the Hosmer-Lemeshow statistic

require(ResourceSelection)

hoslem.test(mobi$mobile, fitted(model))

### compute the Nagelkerke pseudo R square

require(fmsb)

NagelkerkeR2(model)

#### compute all the pseudo R square indicators

require(BaylorEdPsych)

PseudoR2(model)
```

Lesson 13 - Multinomial regression basics

```
news <- read.csv("newspapers.csv")

View(news)

#########
### how to perform the multinomial logistic regression
```

```
#########

#########
### Basic assumptions:

# the dependent variables do not present outliers
# there is no important multicollinearity
##########

### we will determine whether the preferred newspaper is
influenced
### by age and political orientation

### dependent variable: preferred newspaper, with 3
categories
### Daily News, National Politics, Free Tribune

### explainers: age and political opinion (political)
### political is a categorical variable with three
categories:
### Left-wing, Right-wing, Center

### before running the regression, we must set the
reference category (baseline)
### for the categorical variables in the model

## set the baseline

news$newspaper <- relevel(news$newspaper, ref="Free
Tribune")

news$political <- relevel(news$political, ref="Center")

### run the multinomial regression
### using the nnet package, multinom function

require(nnet)

model <- multinom(newspaper~age+political, data = news)

summ <- summary(model)

print(summ)
```

```
### the multinom function only computes the coefficients
and their standard errors
### it does not compute the p values

### we must compute the p values manually

### we compute the z scores first

z <- summ$coefficients/summ$standard.errors

### we generate the p values of the z scores (two-tailed)

pv <- pnorm(abs(z), lower.tail = F) * 2

print(pv)
```

Lesson 14 - Multinomial regression – coefficients

```
news <- read.csv("newspapers.csv")

View(news)

#########
### multinomial logistic regression - compute and interpret
the odds (antilogs of coefficients)
#########

require(nnet)

### set the reference categories and execute the multinom
function

news$newspaper <- relevel(news$newspaper, ref="Free
Tribune")

news$political <- relevel(news$political, ref="Center")

model <- multinom(newspaper~age+political, data = news)

### compute the antilogarithms of the coefficients

expb <- exp(coef(model))
```

```
print(expb)

### compute the confidence intervals for the coefficients

ci <- confint(model, level = 0.95)

print(ci)

### compute the confidence intervals for the antilogarithms

expci <- exp(ci)

print(expci)

### compute the predicted probabilities

pred <- fitted(model)

View(pred)
```

Lesson 15 - Multinomial regression - goodness-of-fit measures

```
news <- read.csv("newspapers.csv")

View(news)

#########
### multinomial logistic regression - goodness-of-fit
measures
#########

require(nnet)

### set the reference categories

news$newspaper <- relevel(news$newspaper, ref="Free
Tribune")

news$political <- relevel(news$political, ref="Center")


### create the null model (without explainers)
```

```
model0 <- multinom(newspaper~1, data = news)

### create our proposed model

model <- multinom(newspaper~age+political, data = news)

### compute the log-likelihoods for both models

LL1 <- logLik(model)
LL0 <- logLik(model0)

### McFadden pseudo R square

mcfadden <- 1 - (LL1 / LL0)

print(mcfadden)

### Cox-Snell pseudo R square

n <- nrow(news)

coxsnell <- 1 - exp((2/n) * (LL0 - LL1))

print(coxsnell)

### Nagelkerke pseudo R square

nagel <- (1 - exp((2/n) * (LL0 - LL1))) / (1 -
exp(LL0)^(2/n))

print(nagel)

### get the deviance

deviance(model)
```

Lesson 16 - Ordinal regression

```
satis <- read.csv("satisfaction.csv")

View(satis)

#########
```

```
### how to perform the ordinal logistic regression
#########

#########
### Basic assumptions:

# the dependent variables do not present outliers
# there is no important multicollinearity
# the condition of proportional odds is met*

### we will only check the assumptions marked with an
asterisk (*)
##########

### we will determine whether the satisfaction level
depends on
### the other variables

### dependent variable: satisfaction with the hotel
services
### 1 - not at all satisfied, 4 - very satisfied

### the explainers are the following:
### customer age
### customer type: pleasure traveller or business traveller
### importance of price: 1 - not important, 2 - somewhat
important, 3 - very important

### N.B. the ordinal variables must be coded numerically
### the nominal variables can be string

### load the package

require(MASS)

## set the baselines (reference categories) for the
categorical explainers

satis$imprice <- relevel(factor(satis$imprice), ref="3")

satis$type <- relevel(satis$type, ref="Business traveler")

model <- polr(factor(satisfaction)~type+age+imprice, data =
satis, method = "logistic")
```

```
summary(model)

### compute the p values for the coefficients

cft <- coef(summary(model))

print(cft)

pv <- pnorm(abs(cft[,"t value"]), lower.tail = F) * 2

print(pv)

### add the p values to the coefficients table

cft <- cbind(cft, "p value" = pv)

print(cft)
```

Lesson 17 - Ordinal regression – coefficients

```
satis <- read.csv("satisfaction.csv")

View(satis)

#########
### ordinal logistic regression - interpreting the
antilogarithms (odds)
#########

### let's run the model again

require(MASS)

### set the baselines (reference categories)

satis$imprice <- relevel(factor(satis$imprice), ref="3")

satis$type <- relevel(satis$type, ref="Business traveler")

model <- polr(factor(satisfaction)~type+age+imprice, data =
satis, method = "logistic")
```

```
### compute the odds (antilogarithms of the coefficients)

odds <- exp(coef(model))

print(odds)

### get the confidence interval for the odds

ci <- exp(confint(model))

print(ci)
```

Lesson 18 - Ordinal regression - goodness-of-fit measures

```
satis <- read.csv("satisfaction.csv")

View(satis)

#########
### ordinal logistic regression - goodness-of-fit measures
#########

### run the model again

require(MASS)

### set the baselines (reference categories)

satis$imprice <- relevel(factor(satis$imprice), ref="3")

satis$type <- relevel(satis$type, ref="Business traveler")

model <- polr(factor(satisfaction)~type+age+imprice, data =
satis, method = "logistic")

### we will compute the goodness-of-fit indicators manually
### based on the log-likelihoods

### first we fit the null model (without independent
variables)

model0 <- polr(factor(satisfaction)~1, data = satis, method
= "logistic")
```

```r
### now we compute the log-likelihood of both null and
proposed model

LL0 <- logLik(model0)
LL1 <- logLik(model)

###### compute the pseudo R squares

### McFadden pseudo R square

mcfadden <- 1 - (LL1 / LL0)

print(mcfadden)

### Cox-Snell pseudo R square

n <- nrow(satis)

coxsnell <- 1 - exp((2/n) * (LL0 - LL1))

print(coxsnell)

### Nagelkerke pseudo R square

nagel <- (1 - exp((2/n) * (LL0 - LL1))) / (1 -
exp(LL0)^(2/n))

print(nagel)

#############

### compute the deviance

deviance(model)

### get the deviance table of the model

require(car)

Anova(model)

### this table displays the statistical significance of
each independent variable
```

Lesson 19 - Ordinal regression - assumption of proportional odds

```
satis <- read.csv("satisfaction.csv")

View(satis)

#########
### ordinal logistic regression - checking the assumption
of proportional odds
#########

### we will use the clm function in the package ordinal

require(ordinal)

model <- clm(factor(satisfaction)~type+age+imprice, data =
satis)

nominal_test(model)

###  nominal_test provides likelihood ratio tests of the
proportional odds assumption
```