Logistic Regression with More Than Two Levels

MSBA7002: Business Statistics

Contents

Contents

1. Car Preference Data, Nominal Logistic Regression

1.1 Prepare data:

```
car <- data.frame(res.unim=c(26, 9, 5, 40, 17, 8), res.im=c(12, 21, 14, 17, 15, 15),
                  res.veim=c(7, 15, 41, 8, 12, 18), sex=c(rep("F", 3), rep("M",3)),
                  age=rep(c("18-23", "24-40", ">40"), 2))
car$age=factor(car$age, levels=c("18-23", "24-40", ">40"))
```

1.2 Fit nomial logistic regression:

```
library(nnet)
options(contrasts=c("contr.treatment", "contr.poly"))
car.mult <- multinom(cbind(res.unim, res.im, res.veim)~sex+age, data=car)
summary(car.mult)
```

glmnet can be used by setting lambda=0. No inference provided though.

1.3 Fit separate binary logistic regression:

```
library(nnet)
options(contrasts=c("contr.treatment", "contr.poly"))
car.bin <- multinom(cbind(res.unim, res.im)~sex+age, data=car)
summary(car.bin)
```

```
library(nnet)
options(contrasts=c("contr.treatment", "contr.poly"))
car.bin <- multinom(cbind(res.unim, res.veim)~sex+age, data=car)
summary(car.bin)
```

This generates different results because π_1 is shared by the two log-likelihood functions.

2. Car Preference Data, Ordinal Logistic Regression

Fit proportional odds logisitic regression:

```
library(MASS)   ### the library containing polr
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
freq <- c(car$res.unim, car$res.im, car$res.veim)
res <- c(rep(c("unim", "im", "veim"), c(6,6,6)))
res <- factor(res, levels=c("unim", "im", "veim"), ordered=T)
car.ord <- data.frame(res=res, sex=rep(car$sex, 3), age=rep(car$age, 3), freq=freq)
car.ord$age=factor(car.ord$age, levels=c("18-23", "24-40", ">40"))
car.polr <- polr(res~sex+age, data=car.ord, weights=freq)
car.polr
```

```
predict(car.polr, car.ord, type="p")
```

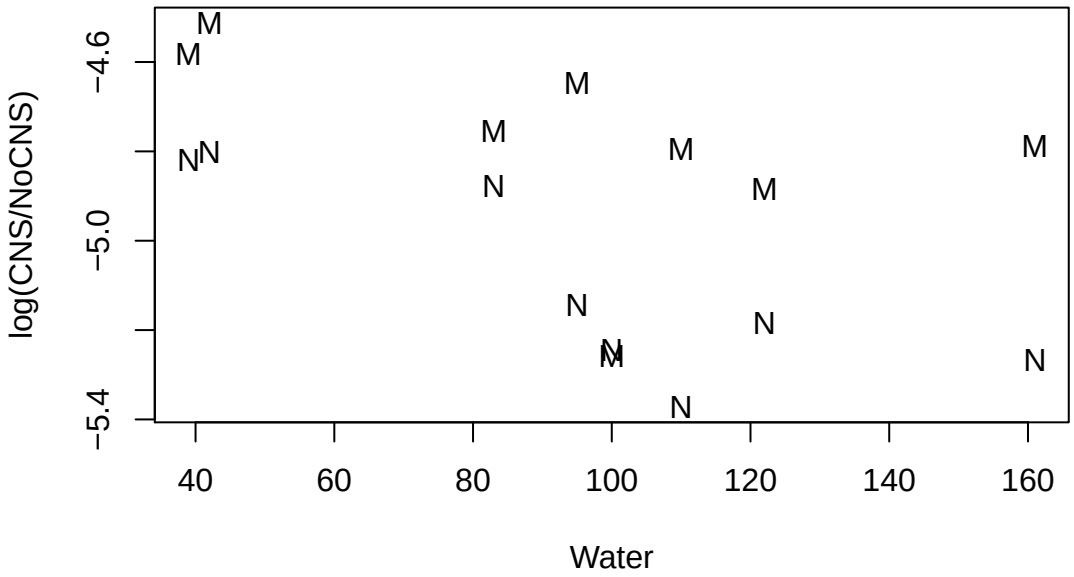
3. Analysis of the CNS data

The data concern live births with deformations of the central nervous system in south Wales.

3.1 Prepare data:

```
cns <- read.csv("cns.TXT", sep=" ", header=T, as.is=T)
cns
names(cns)
```

```
cns$CNS <- cns$An+cns$Sp+cns$Other
plot(log(CNS/NoCNS)~Water, cns, pch=as.character(Work))
```



Note one potential outlier, falls with increasing water hardness, higher for manual workers.

3.2 Fit (binomial) logistic regresion model:

```
binmodw <- glm(cbind(CNS, NoCNS)~Water+Work, cns, family=binomial)
summary(binmodw)
```

Both Water and Work are significant.

How to interpret the coefficients?

3.3 Fit multinomial model with 3 levels:

```
library(nnet)
cmmod <- multinom(cbind(An, Sp, Other)~Water+Work, cns)
summary(cmmod)
### by looking at the coefficients and se, none of the covariates appear to be significant.
nmod <- step(cmmod)
### confirm that the best model is with no covariate
nmod
```

The fitted proportions:

```
cc <- c(0, 0.28963, -0.98083)
names(cc) <- c("An", "Sp", "Other")
exp(cc)/sum(exp(cc))
```

3.4 What happens if we fit a multinomial logit model with 4 categories to begin with?

```
mmod <- multinom(cbind(NoCNS, An, Sp, Other)~Water+Work, cns)
step(mmod)
```

Both are significant! Mis-leading!