26 BE 7024 & 26 PH 7024

Computational Statistics

Autumn, 2014

MB Rao

Homework Sheet No. 4 Due date: October 02, 2014 Maximum points: 30

Multinomial Logistic Regression

The data for this problem has been posted in the ‘Blackboard’ under the name ‘Wheat.’ The data was discussed in the paper:

Martin, C., Harrman, T., Loughin, T., and Oentong, S. (1998). Micropycnometer measurement of single-kernel density of healthy, sprouted, and scab-damaged wheat, *Cereal Chemistry,* 75, 177-180.

Wheat is harvested. Some kernels are healthy (the most desirable), some sprouted (not bad), and the rest scab-infested. It is important to identify each kernel as to its health. Visual inspection is reliable (gold standard) but out of question. The question is what physical characteristics contribute to the state of the kernel. The following characteristics are easy to identify or measurable and the measurements can be mechanized.

class: hrw (hard red winter) or srw (soft red winter)

density

hardness

size

weight

moisture

A sample of 275 kernels yielded data on the characteristics and health (type). I have downloaded the data into R and the following are the top six rows.

> Wheat <- read.delim("clipboard")

> dim(Wheat)

[1] 275 7

> head(Wheat)

class density hardness size weight moisture type

1 hrw 1.349253 60.32952 2.30274 24.6480 12.01538 Healthy

2 hrw 1.287440 56.08972 2.72573 33.2985 12.17396 Healthy

3 hrw 1.233985 43.98743 2.51246 31.7580 11.87949 Healthy

4 hrw 1.336534 53.81704 2.27164 32.7060 12.11407 Healthy

5 hrw 1.259040 44.39327 2.35478 26.0700 12.06487 Healthy

6 hrw 1.300258 48.12066 2.49132 33.2985 12.18577 Healthy

The response variable is ‘type.’

1. How many kernels are of ‘hrw?’ ‘srw?’

> summary(Wheat$class)

hrw srw

1. 2 2 points
2. How many kernels are ‘Healthy?’ ‘Sprout?’ ‘Scab?’

> summary(Wheat$type)

Healthy Scab Sprout

1. 83 96 2 points
2. Get the summary statistics of the data. 3 points

> summary(Wheat)

class density hardness size weight moisture type

hrw:143 Min. :0.7352 Min. :-44.080 Min. :0.5973 Min. : 8.532 Min. : 6.486 Healthy:96

srw:132 1st Qu.:1.1358 1st Qu.: 0.689 1st Qu.:1.8900 1st Qu.:21.982 1st Qu.: 9.540 Scab :83

Median :1.2126 Median : 24.465 Median :2.2303 Median :27.610 Median :11.909 Sprout :96

Mean :1.1885 Mean : 25.564 Mean :2.2047 Mean :27.501 Mean :11.192

3rd Qu.:1.2687 3rd Qu.: 45.606 3rd Qu.:2.5125 3rd Qu.:32.882 3rd Qu.:12.538

Max. :1.6454 Max. :111.934 Max. :4.3100 Max. :46.334 Max. :14.514

1. Postulate the multinomial logistic regression model. 3 points

Pr(Y = 1 / X1, X2,X3,X4,X5,X6) = 

Pr(Y = 2 / X1, X2,X3,X4,X5,X6) = 

Pr(Y = 3 / X1, X2,X3,X4,X5,X6) = ,



where D = 1 + + 

1. Fit the model using the ‘nnet’ package. Write the fitted model. What was the baseline? 5 points

> library(nnet)

> Wheat1<-multinom(formula = type ~ class + density + hardness + size + weight + moisture, data = Wheat)

# weights: 24 (14 variable)

initial value 302.118379

iter 10 value 234.991271

iter 20 value 192.127549

final value 192.112352

converged

> summary(What1)

Error in summary(What1) : object 'What1' not found

> summary(Wheat1)

Call:

multinom(formula = type ~ class + density + hardness + size +

weight + moisture, data = Wheat)

Coefficients:

(Intercept) classsrw density hardness size weight moisture

Scab 30.54650 -0.6481277 -21.59715 -0.01590741 1.0691139 -0.2896482 0.10956505

Sprout 19.16857 -0.2247384 -15.11667 -0.02102047 0.8756135 -0.0473169 -0.04299695

Std. Errors:

(Intercept) classsrw density hardness size weight moisture

Scab 4.289865 0.6630948 3.116174 0.010274587 0.7722862 0.06170252 0.1548407

Sprout 3.767214 0.5009199 2.764306 0.008105748 0.5409317 0.03697493 0.1127188

Residual Deviance: 384.2247

AIC: 412.2247

The category Healthy is the baseline.

Perform the ‘Anova’ command from the ‘car’ package to identify the significant predictors. 4 points

> library(car)

> Anova(Wheat1)

Analysis of Deviance Table (Type II tests)

Response: type

LR Chisq Df Pr(>Chisq)

class 0.964 2 0.6175

density 90.555 2 < 2.2e-16 \*\*\*

hardness 7.074 2 0.0291 \*

size 3.211 2 0.2008

weight 28.230 2 7.411e-07 \*\*\*

moisture 1.193 2 0.5506

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 The variables density and weight are the most significant

1. Change the baseline to whatever you like and refit the model. 4 points

> head(Wheat$type)

[1] Healthy Healthy Healthy Healthy Healthy Healthy

Levels: Healthy Scab Sprout

> Wheat$type1 <- relevel(Wheat$type, ref = "Scab")

> head(Wheat$type1)

[1] Healthy Healthy Healthy Healthy Healthy Healthy

Levels: Scab Healthy Sprout

> Wheat2<-multinom(formula = type1 ~ class + density + hardness + size + weight + moisture, data = Wheat)

# weights: 24 (14 variable)

initial value 302.118379

iter 10 value 224.607285

iter 20 value 192.123211

final value 192.112352

converged

> summary(Wheat2)

Call:

multinom(formula = type1 ~ class + density + hardness + size +

weight + moisture, data = Wheat)

Coefficients:

(Intercept) classsrw density hardness size weight moisture

Healthy -30.54719 0.6480837 21.597998 0.015906072 -1.0692962 0.2896538 -0.1095643

Sprout -11.37844 0.4233489 6.480898 -0.005113569 -0.1935887 0.2423401 -0.1525601

Std. Errors:

(Intercept) classsrw density hardness size weight moisture

Healthy 4.289936 0.6630981 3.116251 0.010274660 0.7722932 0.06170339 0.1548419

Sprout 2.641890 0.6040995 1.852922 0.008597455 0.6774480 0.05581175 0.1409534

Residual Deviance: 384.2247

AIC: 412.2247

1. Use the command ‘predict’ to predict the type of each kernel. Present the top six rows of the consolidated data along with the predicted type. Calculate the misclassification rate. 7 points

> Wheat3 <- predict(Wheat2, newdata = Wheat, type = "class")

> head(Wheat3)

[1] Healthy Healthy Healthy Healthy Healthy Healthy

Levels: Scab Healthy Sprout

> Wheat4 <- data.frame(Wheat, ClassPred = Wheat3)

> head(Wheat4)

class density hardness size weight moisture type type1 ClassPred

1 hrw 1.349253 60.32952 2.30274 24.6480 12.01538 Healthy Healthy Healthy

2 hrw 1.287440 56.08972 2.72573 33.2985 12.17396 Healthy Healthy Healthy

3 hrw 1.233985 43.98743 2.51246 31.7580 11.87949 Healthy Healthy Healthy

4 hrw 1.336534 53.81704 2.27164 32.7060 12.11407 Healthy Healthy Healthy

5 hrw 1.259040 44.39327 2.35478 26.0700 12.06487 Healthy Healthy Healthy

6 hrw 1.300258 48.12066 2.49132 33.2985 12.18577 Healthy Healthy Healthy

> Wheat5 <- ifelse(Wheat$type == Wheat4$ClassPred, 1, 0)

> Wheat5

[1] 1 1 1 1 1 1 1 1 1 1 0 1 1 0 0 0 0 0 0 0 1 0 1 0 1 0 1 0 1 1 1 0 1 0 0 0 0 1 1 0 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 0 1 1 1 1 1 0 1 1 1 0 0 0

[80] 0 0 1 0 1 0 1 0 1 0 0 0 0 0 1 1 1 1 0 0 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0 1 0 0 1 0 1 0 1 1 1 1 0 0 1 1 1 1 1 1 1 0 0 0 0 1 0 0 0 1 1 1 1 0 1

[159] 1 1 0 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 0 0 1 1 1 1 0 1 0 1 1 1 1 1 1 1 1 0 1 0 0 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 0 1 1

[238] 1 0 0 1 1 1 0 1 1 1 1 1 1 1 0 0 1 1 1 1 0 1 0 1 1 1 1 1 1 0 1 1 0 1 0 1 1 1