26 BE 7082 + 26 PH 7028 + 20 BME 7082

Autumn 2020

MB Rao

Mid-Term Project Due Date: October 29, 2020 Maximum Points: 40

Theme: Learn how to fit the multinomial logistic regression model to the data. This is an elaborate project. Models for categorical response variables are always like that.

What does a high school student do after graduation? There are three possibilities: general (seek a job); enroll into an academic institution; enroll into a vocational school. What are the factors shaping the choice? A researcher collected data on 200 high school students and posted the data on the internet as a .dta file. Here is its URL.

<https://stats.idre.ucla..edu/stat/data/hsbdemo.dta>

Information was gathered on 12 variables.

1. female (categorical): male or female

2. ses (categorical) (socioeconomic status): low; middle; high

3. schtyp (categorical) (school type): High school is either public or private

4. prog (program chosen); general; academic; vocation (This is our response variable, which is nominal)

High school students take State Board exams on the following.

5. read (reading score) – numeric

6. write (writing score) – numeric

7. math (math score) – numeric

8. science (science score) – numeric

9. socst (social studies score) – numeric

10. honors (honors program) – not enrolled; enrolled

11. award (Number of awards won)

12. cid (I don’t know what this is.)

Download and activate the package ‘foreign.’

Download the data. Use the following R code.

MB <- read.dta(<https://stats.idre.ucla.edu/stat/data/hsbdemo.dta>)

*1. What is the dimension of the data? 2 points*

**200 rows and 13 columns.**

> SF <- read.dta("https://stats.idre.ucla.edu/stat/data/hsbdemo.dta")

> dim(SF)

[1] 200 13

*2. Show the top six rows of the data. 2 points*

**> head(SF)**

id female ses schtyp prog read write math science socst honors awards cid

1 45 female low public vocation 34 35 41 29 26 not enrolled 0 1

2 108 male middle public general 34 33 41 36 36 not enrolled 0 1

3 15 male high public vocation 39 39 44 26 42 not enrolled 0 1

4 67 male low public vocation 37 37 42 33 32 not enrolled 0 1

5 153 male middle public vocation 39 31 40 39 51 not enrolled 0 1

6 51 female high public general 42 36 42 31 39 not enrolled 0 1

*3. The first column is id. Create a new folder eliminating the first column. Call this new folder MB1. 2 poins*

**> SF1<-subset(SF, select=-c(id))**

*4. When we want to analyze data for a research article, journals usually request piecemeal analyses. Examine how the response variable is associated with each of the predictors. Each analysis results in either with a contingency table or data in the anova format.*

*a. Cross-tabulate female and prog. Conduct a chi-squared test of independence.*

**> chisq.test(SF1$female, SF1$prog)**

Pearson's Chi-squared test

data: SF1$female and SF1$prog

X-squared = 0.052809, df = 2, p-value = 0.9739

*b. Cross-tabulate ses and prog. Conduct a chi-squared test of independence.*

**> chisq.test(SF1$ses, SF1$prog)**

Pearson's Chi-squared test

data: SF1$ses and SF1$prog

X-squared = 16.604, df = 4, p-value = 0.002307

*c. Cross-tabulate schtype and prog. Conduct a chi-squared test of independence.*

**> chisq.test(SF1$schtyp, SF1$prog)**

Pearson's Chi-squared test

data: SF1$schtyp and SF1$prog

X-squared = 9.2687, df = 2, p-value = 0.009712

*d. Cross-tabulate honors and prog. Conduct a chi-squared test of independence.*

**> chisq.test(SF1$honors, SF1$prog)**

Pearson's Chi-squared test

data: SF1$honors and SF1$prog

X-squared = 15.413, df = 2, p-value = 0.00045

*e. Examine significant differences between the levels of ‘prog’ with respect to ‘read.’ (READ <- anova(lm(read ~ prog, data = MB1) )*

**> READ<-anova(lm(read~prog, data=SF1))**

**> READ**

Analysis of Variance Table

Response: read

Df Sum Sq Mean Sq F value Pr(>F)

prog 2 3716.9 1858.43 21.282 4.283e-09 \*\*\*

Residuals 197 17202.6 87.32

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

*f. Examine significant differences between the levels of ‘prog’ with respect to ‘write.’*

**> READ1<-anova(lm(write~prog, data=SF1))**

**> READ1**

Analysis of Variance Table

Response: write

Df Sum Sq Mean Sq F value Pr(>F)

prog 2 3175.7 1587.85 21.275 4.31e-09 \*\*\*

Residuals 197 14703.2 74.64

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

*g. Examine significant differences between the levels of ‘prog’ with respect to ‘math.’*

**> READ2<-anova(lm(math~prog, data=SF1))**

**> READ2**

**Analysis of Variance Table**

Response: math

Df Sum Sq Mean Sq F value Pr(>F)

prog 2 4002.1 2001.05 29.279 7.364e-12 \*\*\*

Residuals 197 13463.7 68.34

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

*h. Examine significant differences between the levels of ‘prog’ with respect to ‘science.’*

**> READ3<-anova(lm(science~prog, data=SF1))**

**> READ3**

Analysis of Variance Table

Response: science

Df Sum Sq Mean Sq F value Pr(>F)

prog 2 1487 743.50 8.128 0.0004057 \*\*\*

Residuals 197 18021 91.47

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

*i. Examine significant differences between the levels of ‘prog’ with respect to ‘write.’*

**> READ4<-anova(lm(write~prog, data=SF1))**

**> READ4**

Analysis of Variance Table

Response: write

Df Sum Sq Mean Sq F value Pr(>F)

prog 2 3175.7 1587.85 21.275 4.31e-09 \*\*\*

Residuals 197 14703.2 74.64

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

*j. Examine significant differences between the levels of ‘prog’ with respect to ‘socst.’*

**> READ5<-anova(lm(socst~prog, data=SF1))**

**> READ5**

Analysis of Variance Table

Response: socst

Df Sum Sq Mean Sq F value Pr(>F)

prog 2 4806.2 2403.08 26.112 8.728e-11 \*\*\*

Residuals 197 18130.0 92.03

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

*k. Examine significant differences between the levels of ‘prog’ with respect to ‘awards.’ 11 points*

**> READ6<-anova(lm(awards~prog, data=SF1))**

**> READ6**

Analysis of Variance Table

Response: awards

Df Sum Sq Mean Sq F value Pr(>F)

prog 2 81.54 40.769 13.927 2.201e-06 \*\*\*

Residuals 197 576.68 2.927

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

*5. Fit a multinomial logistic regression model with ‘prog’ as response variable and the rest predictors.*

**> SF2<-vglm(prog~., data=SF1, family=multinomial)**

*a. Show the summary.*

**> summary(SF2)**

Call:

vglm(formula = prog ~ ., family = multinomial, data = SF1)

Pearson residuals:

Min 1Q Median 3Q

log(mu[,1]/mu[,3]) -3.649 -0.4554 -0.2036 -0.09352

log(mu[,2]/mu[,3]) -5.289 -0.5235 0.1952 0.51488

Max

log(mu[,1]/mu[,3]) 2.826

log(mu[,2]/mu[,3]) 3.638

Coefficients:

Estimate Std. Error z value

(Intercept):1 -5.091577 10.092530 -0.504

(Intercept):2 -17.254605 10.654713 -1.619

femalefemale:1 -0.254025 0.524722 -0.484

femalefemale:2 -0.430391 0.501223 -0.859

sesmiddle:1 -1.493127 0.560450 -2.664

sesmiddle:2 -1.252346 0.575752 -2.175

seshigh:1 -0.939230 0.719899 -1.305

seshigh:2 0.042767 0.692577 0.062

schtypprivate:1 1.332254 0.891866 1.494

schtypprivate:2 1.911763 0.831166 2.300

read:1 0.004605 0.076248 0.060

read:2 0.091565 0.078626 1.165

write:1 0.001252 0.080017 0.016

write:2 0.099725 0.085260 1.170

math:1 0.027699 0.077026 0.360

math:2 0.170929 0.080063 2.135

science:1 0.047457 0.076202 0.623

science:2 -0.012420 0.077765 -0.160

socst:1 0.045635 0.028347 1.610

socst:2 0.068392 0.029106 2.350

honorsenrolled:1 -1.661949 1.110051 -1.497

honorsenrolled:2 -1.097246 0.973871 -1.127

awards:1 0.368446 0.409434 0.900

awards:2 0.121404 0.382794 0.317

cid:1 -0.044352 0.392512 -0.113

cid:2 -0.286804 0.401467 -0.714

Pr(>|z|)

(Intercept):1 0.61392

(Intercept):2 0.10535

femalefemale:1 0.62831

femalefemale:2 0.39052

sesmiddle:1 0.00772 \*\*

sesmiddle:2 0.02962 \*

seshigh:1 0.19201

seshigh:2 0.95076

schtypprivate:1 0.13523

schtypprivate:2 0.02144 \*

read:1 0.95185

read:2 0.24420

write:1 0.98752

write:2 0.24214

math:1 0.71914

math:2 0.03277 \*

science:1 0.53343

science:2 0.87311

socst:1 0.10743

socst:2 0.01878 \*

honorsenrolled:1 0.13435

honorsenrolled:2 0.25988

awards:1 0.36818

awards:2 0.75113

cid:1 0.91003

cid:2 0.47499

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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

log(mu[,2]/mu[,3])

Residual deviance: 308.1006 on 374 degrees of freedom

Log-likelihood: -154.0503 on 374 degrees of freedom

Number of Fisher scoring iterations: 5

No Hauck-Donner effect found in any of the estimates

Reference group is level 3 of the response

*b. Identify the baseline.*

**Level 3 is the baseline.**

**> levels(SF1$prog)[3]**

**[1] "vocation"**

*c. Examine goodness-of-fit.*

**> pchisq(308.1006, 374, lower.tail=F)**

**[1] 0.994503 >> 0.05 fit is excellent.**

*d. Do soft prediction.*

**soft<-round(predict(SF2, newdata=SF1, type="response"), 3)**

**> head(soft)**

general academic vocation

1 0.176 0.041 0.783

2 0.122 0.029 0.849

3 0.101 0.457 0.442

4 0.283 0.130 0.587

5 0.228 0.071 0.701

6 0.121 0.231 0.648

*e. Do hard prediction.*

**> Hard <- rep(0,200)**

**> for (i in 1:200)**

**+ (**

**+ Hard[i] <- ifelse (soft[i,1] == max (soft[i, ]), "general", ifelse (soft[i,2] == max(soft[i, ]), "academic", "vocation"))**

**+ )**

**> Hard**

[1] "vocation" "vocation" "academic" "vocation" "vocation" "vocation" "vocation" "vocation" "vocation" "vocation"

[11] "vocation" "vocation" "vocation" "vocation" "vocation" "vocation" "general" "vocation" "vocation" "academic"

[21] "vocation" "vocation" "general" "vocation" "general" "academic" "academic" "vocation" "academic" "academic"

[31] "vocation" "vocation" "vocation" "academic" "vocation" "vocation" "vocation" "academic" "general" "academic"

[41] "vocation" "academic" "academic" "vocation" "academic" "general" "vocation" "general" "academic" "academic"

[51] "vocation" "vocation" "general" "academic" "academic" "general" "academic" "general" "academic" "vocation"

[61] "vocation" "vocation" "academic" "vocation" "academic" "vocation" "general" "academic" "academic" "vocation"

[71] "academic" "vocation" "academic" "academic" "academic" "academic" "academic" "vocation" "academic" "general"

[81] "general" "vocation" "academic" "general" "academic" "academic" "academic" "academic" "general" "vocation"

[91] "vocation" "academic" "academic" "academic" "vocation" "academic" "academic" "academic" "academic" "academic"

[101] "general" "vocation" "general" "vocation" "vocation" "academic" "academic" "academic" "academic" "academic"

[111] "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "vocation"

[121] "vocation" "academic" "academic" "academic" "academic" "academic" "general" "general" "academic" "academic"

[131] "academic" "academic" "general" "academic" "academic" "academic" "academic" "academic" "general" "academic"

[141] "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic"

[151] "academic" "academic" "general" "academic" "academic" "academic" "academic" "academic" "general" "academic"

[161] "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic"

[171] "general" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic"

[181] "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic"

[191] "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic"

*f. Lay out the confusion matrix.*

**> #confusion matrix**

**> confusion <- data.frame(observed = SF1$prog, predicted = Hard)**

**> confusion1<-table(confusion$observed, confusion$predicted)**

**> confusion1**

academic general vocation

general 23 11 11

academic 89 8 8

vocation 16 4 30

*g. Calculate the misclassification rate. 7 points*

**misclassification rate = [(23+11)+(8+8)+(16+4)]/200 = 70/200 = 35%**

*6. Fit a multinomial logistic regression model with ‘prog’ as response variable and ‘ses’ and ‘write’ as predictors.*

**SF3<-vglm(prog~ ses+write, data=SF1, family=multinomial)**

*a. Show the summary.*

**> summary(SF3)**

Call:

vglm(formula = prog ~ ses + write, family = multinomial, data = SF1)

Pearson residuals:

Min 1Q Median 3Q

log(mu[,1]/mu[,3]) -3.260 -0.4332 -0.2428 -0.1854

log(mu[,2]/mu[,3]) -4.675 -0.6697 0.3659 0.7600

Max

log(mu[,1]/mu[,3]) 2.592

log(mu[,2]/mu[,3]) 2.149

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept):1 -2.36601 1.17423 -2.015 0.0439 \*

(Intercept):2 -5.21820 1.16351 -4.485 7.30e-06 \*\*\*

sesmiddle:1 -0.82468 0.49012 -1.683 0.0924 .

sesmiddle:2 -0.29139 0.47637 -0.612 0.5407

seshigh:1 -0.18016 0.64842 -0.278 0.7811

seshigh:2 0.98267 0.59553 1.650 0.0989 .

write:1 0.05567 0.02333 2.386 0.0170 \*

write:2 0.11360 0.02222 5.113 3.17e-07 \*\*\*

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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

log(mu[,2]/mu[,3])

Residual deviance: 359.9635 on 392 degrees of freedom

Log-likelihood: -179.9817 on 392 degrees of freedom

Number of Fisher scoring iterations: 4

No Hauck-Donner effect found in any of the estimates

Reference group is level 3 of the response

*b. Identify the baseline.*

**Level 3 is the baseline.**

**> levels(SF1$prog)[3]**

**[1] "vocation"**

*c. Examine goodness-of-fit.*

**> pchisq(359.9635, 392, lower.tail=F)**

**[1] 0.8755298 >> 0.05 fit is excellent**

*d. Do soft prediction.*

**soft2<- round(predict(SF3, newdata=SF1, type="response"), 3)**

**> head(soft2)**

general academic vocation

1 0.338 0.148 0.513

2 0.181 0.120 0.699

3 0.237 0.419 0.345

4 0.351 0.173 0.476

5 0.169 0.100 0.731

6 0.238 0.353 0.409

*e. Do hard prediction.*

**> Hard2 <- rep(0,200)**

**> for (i in 1:200)**

**+ (**

**+ Hard2[i] <- ifelse (soft2[i,1] == max (soft2[i, ]), "general", ifelse (soft2[i,2] == max(soft2[i, ]), "academic", "vocation"))**

**+ )**

**> Hard2**

[1] "vocation" "vocation" "academic" "vocation" "vocation" "vocation" "vocation" "vocation" "vocation" "vocation"

[11] "general" "vocation" "vocation" "vocation" "vocation" "general" "general" "vocation" "vocation" "academic"

[21] "vocation" "vocation" "general" "general" "vocation" "academic" "general" "vocation" "vocation" "academic"

[31] "vocation" "vocation" "vocation" "general" "vocation" "vocation" "vocation" "academic" "vocation" "vocation"

[41] "vocation" "academic" "academic" "academic" "general" "academic" "vocation" "academic" "vocation" "academic"

[51] "vocation" "vocation" "general" "academic" "academic" "general" "academic" "academic" "academic" "vocation"

[61] "academic" "vocation" "academic" "academic" "academic" "vocation" "academic" "academic" "academic" "academic"

[71] "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "vocation" "vocation"

[81] "general" "academic" "academic" "general" "academic" "general" "academic" "vocation" "academic" "vocation"

[91] "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "vocation" "vocation"

[101] "academic" "academic" "academic" "vocation" "academic" "academic" "academic" "academic" "academic" "academic"

[111] "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic"

[121] "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic"

[131] "academic" "academic" "general" "academic" "academic" "academic" "academic" "academic" "general" "academic"

[141] "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic"

[151] "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic"

[161] "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic"

[171] "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic"

[181] "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic"

[191] "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic" "academic"

>

*f. Lay out the confusion matrix.*

**> confusion2<-data.frame(observed=SF1$prog, predicted = Hard2)**

**> confusion3<-table(confusion2$observed, confusion2$predicted)**

**> confusion3**

academic general vocation

general 27 7 11

academic 92 4 9

vocation 23 4 23

*g. Calculate the misclassification rate.*

**Misclassification rate = [(27+11)+(4+9)+(23+4)]/200 = 78/200 = 39%**

*h. Write the prediction equation.*

**Pr(general) <- exp(-2.36601-0.82468\*sesmiddle-0.18016\*seshigh+0.05567\*write)/D**

**Pr(academic) <- exp(-5.21820-0.29139\*sesmiddle+0.98267\*seshigh+0.11360\*write)/D**

**Pr(vocation)=1/D**

**D = 1+ [exp(-2.36601-0.82468\*sesmiddle-0.18016\*seshigh+0.05567\*write)/] + [exp(-5.21820-0.29139\*sesmiddle+0.98267\*seshigh+0.11360\*write)]**

**Here, SES for students belonging to**

**High SES <- sesmiddle = 0, seshigh=1**

**Middle SES <- sesmiddle = 1, seshigh=0**

**Low SES<- sesmiddle = 0, seshigh=0**

*i. Interpret the output.*

**1) Students’ writing score is a strong indicator of their program of choice.**

**2) Students are more likely (Pr >0.5) to to pursue higher studies in academics with high writing scores (>60).**

**3) Students with higher socioeconomic status are more likely to pursue higher studies compared to students belonging to low and medium socioeconomic status.**

**4) Vocational programs are more likely (Pr>0.5) to be opted by students with low writing scores (<40).**

**5) Students from higher socioeconomic background are less likely to opt for vocational programs compared to students belonging to low and medium socioeconomic background.**

**6) Students belonging to lower socioeconomic status are more likely (Pr ~0.45) to opt for jobs when they have low to medium writing scores(40~60) compared to those belonging to middle and high socioeconomic status (Pr <0.25).**

**7) In my opinion, there seems to be a relation between socioeconomic status and writing scores and overall pursuit of academic studies.**

**8) This can be explained by the availability of resources and financial freedom of students that belong to higher and high to middle socioeconomic status.**

*j. Present the model graphically. Take x = write and y = probability. I would be seeing 9 curves. 9 + 6 points*

**I generated curves with “ggplot” package.**

SF4 <- read.dta("https://stats.idre.ucla.edu/stat/data/hsbdemo.dta")

with(SF4, table(ses,prog))

with(SF4, do.call(rbind, tapply(write, prog, function(x) c(M=mean(x), SD = sd(x)))))

SF4$prog2<-relevel(SF4$prog, ref="academic")

test<-multinom(prog2 ~ ses+write, data=SF4)

dses <- data.frame(ses = c("low", "middle", "high"), write = mean(SF4$write))

dwrite<- data.frame(ses=rep(c("low", "middle", "high"), each = 41), write = rep(c(30:70),3))

pp.write<- cbind(dwrite, predict(test, new=dwrite, type="probs", se=TRUE))

lpp <- melt (pp.write, id.vars = c("ses", "write"), value.name = "probability")

head(lpp)

ggplot(lpp, aes(x=write, y=probability, colour=ses)) + geom\_line() + facet\_grid(variable ~., scales="free")

