# Homework 4

Statistical Modelling with R

Assignment by: Tom Wiesing, Nzaramba Julius, Ashishsingh Raja Rai

> library(car)

> library(MASS)

> library(xtable)

> library(epitools)

> library(nnet)

> library(foreign)

> library(ggplot2)

> library(reshape2)

**Problem 1:**

> student <- read.dta("<http://www.ats.ucla.edu/stat/data/hsbdemo.dta>")

> ex1 <- xtabs(~ ses + prog, data=student)

> ex1

prog

ses general academic vocation

low 16 19 12

middle 20 44 31

high 9 42 7

>

a) The academic program was chosen by the largest fraction with high ses

b) 34.04255% of low income students selected the general program

> ex1 / rowSums(ex1, na.rm = T)

prog

ses general academic vocation

low 0.3404255 0.4042553 0.2553191

middle 0.2105263 0.4631579 0.3263158

high 0.1551724 0.7241379 0.1206897

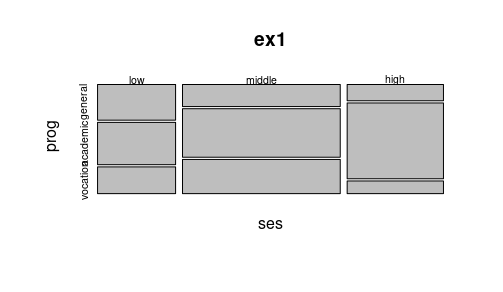
c) In the academic program, there are more students with middle social economic status.

d) The least frequent combination of variables is vocation / high ses

**Problem 2:**

a)

> mosaicplot(ex1)



b)

> # Make a copy of the data and add the aggregated attribute.

> studentcopy <- student

> studentcopy$academicProgram <- NA

> studentcopy[studentcopy$prog == "academic",]$academicProgram <- "yes"

> studentcopy[studentcopy$prog != "academic",]$academicProgram <- "no"

>

> # compute the odds ratio

> ex2b <- xtabs(~ ses + academicProgram, data=studentcopy)

> oddsratio(ex2b)

$data

academicProgram

ses no yes Total

low 28 19 47

middle 51 44 95

high 16 42 58

Total 95 105 200

$measure

odds ratio with 95% C.I.

ses estimate lower upper

low 1.000000 NA NA

middle 1.267139 0.6238113 2.610207

high 3.798125 1.6900452 8.852748

$p.value

two-sided

ses midp.exact fisher.exact chi.square

low NA NA NA

middle 0.513992166 0.591185157 0.5061643631

high 0.001118016 0.001386481 0.0009553094

$correction

[1] FALSE

attr(,"method")

[1] "median-unbiased estimate & mid-p exact CI"

> # low / middle = 19 / 44 = 0.43

> 19 / 44

[1] 0.4318182

> # low / high = 19 /42 = 0.45

> 19 / 42

[1] 0.452381

>

**Problem 3:**

> ex3 <- chisq.test(ex1)

> ex3

Pearson's Chi-squared test

data: ex1

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

a) As p <0.05, the chi^2 test statistic is significant.

b) The expected frequencies are higher in low/academic, middle/general, middle/academic, high/general, high/vocation

> ex3$expected > ex1

prog

ses general academic vocation

low FALSE TRUE FALSE

middle TRUE TRUE FALSE

high TRUE FALSE TRUE

**Problem 4:**

> # split data into male and female

> stud\_female <- student[student$female == "female", ]

> stud\_male <- student[student$female == "male", ]

a)

> ex4a\_f <- xtabs(~ ses + prog, data=stud\_female)

> ex4a\_m <- xtabs(~ ses + prog, data=stud\_male)

>

> sq4a\_f <- chisq.test(ex4a\_f)

> sq4a\_f # females : not significant

Pearson's Chi-squared test

data: ex4a\_f

X-squared = 7.473, df = 4, p-value = 0.1129

> sq4a\_m <- chisq.test(ex4a\_m)

Warning message:

In chisq.test(ex4a\_m) : Chi-squared approximation may be incorrect

> sq4a\_m # males: significant

Pearson's Chi-squared test

data: ex4a\_m

X-squared = 11.557, df = 4, p-value = 0.02097

> x

b)

> sq4a\_f$expected > ex4a\_f # females: low/academic, middle/general, middle/academic, high/general, high/vocation

prog

ses general academic vocation

low FALSE TRUE FALSE

middle TRUE TRUE FALSE

high TRUE FALSE TRUE

> sq4a\_m$expected > ex4a\_m # males: low/academic, middle/general, middle/academic, high/general, high/vocation

prog

ses general academic vocation

low FALSE TRUE FALSE

middle TRUE TRUE FALSE

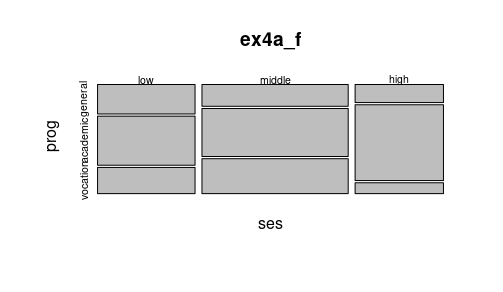
high TRUE FALSE TRUE

>

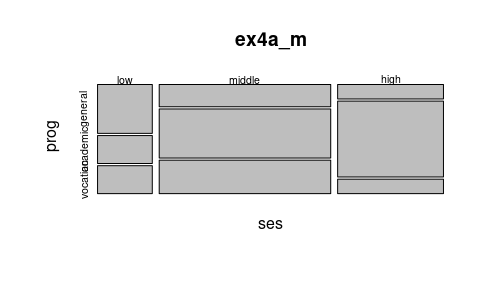
c) 4a: yes, 4b: no

d) the low/general ratio is different

> mosaicplot(ex4a\_f)



> mosaicplot(ex4a\_m)



**Problem 5:**

> ex5 <- multinom(prog ~ female + ses + schtyp + read + write + math + science + honors + awards, data=student, trace=FALSE)

a) The AIC is 358.8871

> AIC(ex5)

[1] 358.8871

b) Computing the p values and test statistics

> summary(ex5, cor=FALSE, Wald=TRUE)

Call:

multinom(formula = prog ~ female + ses + schtyp + read + write +

math + science + honors + awards, data = student, trace = FALSE)

Coefficients:

(Intercept) femalefemale sesmiddle seshigh schtypprivate

academic -5.498379 -0.1522507 0.3751645 1.0802062 0.5372474

vocation 4.163618 0.2804404 1.3235922 0.7959664 -1.2758046

read write math science honorsenrolled

academic 0.052955458 0.06429399 0.10006643 -0.10199019 0.5743597

vocation -0.009609139 -0.02166161 -0.02995725 -0.03800852 1.9991952

awards

academic -0.2703505

vocation -0.3525181

Std. Errors:

(Intercept) femalefemale sesmiddle seshigh schtypprivate read

academic 2.347466 0.4533749 0.5061641 0.5772063 0.5550686 0.02919946

vocation 2.455693 0.5179411 0.5388177 0.7084789 0.8817585 0.03322165

write math science honorsenrolled awards

academic 0.04927104 0.03480492 0.03090596 0.8644712 0.2936491

vocation 0.05054331 0.03756352 0.03224286 1.1038196 0.3856715

Value/SE (Wald statistics):

(Intercept) femalefemale sesmiddle seshigh schtypprivate read

academic -2.342262 -0.3358164 0.7411914 1.871439 0.9678937 1.8135768

vocation 1.695496 0.5414523 2.4564750 1.123486 -1.4468866 -0.2892433

write math science honorsenrolled awards

academic 1.3049044 2.8750657 -3.300016 0.6644058 -0.9206585

vocation -0.4285752 -0.7975091 -1.178820 1.8111612 -0.9140370

Residual Deviance: 314.8871

AIC: 358.8871

> z <- summary(ex5, cor=FALSE, Wald=TRUE)$Wald.ratios

> p <- (1 - pnorm(abs(z), 0, 1))\*2

> p

(Intercept) femalefemale sesmiddle seshigh schtypprivate read

academic 0.01916726 0.7370093 0.45857742 0.06128429 0.3330974 0.0697429

vocation 0.08998133 0.5881959 0.01403076 0.26123099 0.1479287 0.7723952

write math science honorsenrolled awards

academic 0.1919254 0.004039434 0.0009667921 0.50643061 0.3572288

vocation 0.6682324 0.425155392 0.2384699548 0.07011591 0.3606974

> p < 0.05

(Intercept) femalefemale sesmiddle seshigh schtypprivate read write

academic TRUE FALSE FALSE FALSE FALSE FALSE FALSE

vocation FALSE FALSE TRUE FALSE FALSE FALSE FALSE

math science honorsenrolled awards

academic TRUE TRUE FALSE FALSE

vocation FALSE FALSE FALSE FALSE

> # signigicant: academic / intercept, academic/math, academic/science, vocation/sesmiddle

>

**Problem 6:**

> ex6 <- stepAIC(ex5, direction="backward")

Start: AIC=358.89

prog ~ female + ses + schtyp + read + write + math + science +

honors + awards

Df AIC

- female 2 355.67

- awards 2 356.01

- write 2 358.17

- honors 2 358.34

<none> 358.89

- read 2 359.95

- schtyp 2 361.51

- ses 4 362.44

- science 2 367.39

- math 2 370.84

Step: AIC=355.67

prog ~ ses + schtyp + read + write + math + science + honors +

awards

Df AIC

- awards 2 352.76

- write 2 354.33

- honors 2 354.93

<none> 355.67

- read 2 357.11

- schtyp 2 358.13

- ses 4 359.29

- science 2 364.14

- math 2 367.96

Step: AIC=352.76

prog ~ ses + schtyp + read + write + math + science + honors

Df AIC

- honors 2 351.68

<none> 352.76

- write 2 353.87

- read 2 354.15

- schtyp 2 355.13

- ses 4 355.90

- science 2 360.93

- math 2 365.06

Step: AIC=351.68

prog ~ ses + schtyp + read + write + math + science

Df AIC

- write 2 350.99

<none> 351.68

- read 2 352.60

- ses 4 353.85

- schtyp 2 355.13

- science 2 359.88

- math 2 363.23

Step: AIC=350.99

prog ~ ses + schtyp + read + math + science

Df AIC

<none> 350.99

- ses 4 353.66

- read 2 354.20

- schtyp 2 355.73

- science 2 357.99

- math 2 368.31

>

a) ses + schtyp + read + math + science

b) The BIC is 397.1684

> BIC(ex6)

[1] 397.1684

c) -161.496

> logLik(ex6)

'log Lik.' -161.496 (df=14)

>

**Problem 7:**

We need to predict for high/private, middle/private, low/private, high/public, middle/public, low/public

> ex7data <- expand.grid(ses = c("low", "middle", "high"), schtyp=c("public","private"), read=mean(student$read), math=mean(student$math), science = mean(student$science))

> ex7data

ses schtyp read math science

1 low public 52.23 52.645 51.85

2 middle public 52.23 52.645 51.85

3 high public 52.23 52.645 51.85

4 low private 52.23 52.645 51.85

5 middle private 52.23 52.645 51.85

6 high private 52.23 52.645 51.85

> predict(ex6, ex7data, "probs")

general academic vocation

1 0.3597989 0.4821528 0.15804831

2 0.2333606 0.4319957 0.33464371

3 0.1766363 0.6668812 0.15648254

4 0.2801551 0.6897408 0.03010409

5 0.2104441 0.7157334 0.07382253

6 0.1226533 0.8507663 0.02658044

>

**Problem 8:**

> ex8data <- expand.grid(ses = c("low", "middle", "high"), schtyp=c("public","private"), read=mean(student$read), math=mean(student$math), science = seq(30, 80, 1))

> ex8data

ses schtyp read math science

1 low public 52.23 52.645 30

2 middle public 52.23 52.645 30

3 high public 52.23 52.645 30

4 low private 52.23 52.645 30

5 middle private 52.23 52.645 30

6 high private 52.23 52.645 30

7 low public 52.23 52.645 31

8 middle public 52.23 52.645 31

9 high public 52.23 52.645 31

10 low private 52.23 52.645 31

11 middle private 52.23 52.645 31

12 high private 52.23 52.645 31

13 low public 52.23 52.645 32

14 middle public 52.23 52.645 32

15 high public 52.23 52.645 32

16 low private 52.23 52.645 32

17 middle private 52.23 52.645 32

18 high private 52.23 52.645 32

19 low public 52.23 52.645 33

20 middle public 52.23 52.645 33

21 high public 52.23 52.645 33

22 low private 52.23 52.645 33

23 middle private 52.23 52.645 33

24 high private 52.23 52.645 33

25 low public 52.23 52.645 34

26 middle public 52.23 52.645 34

27 high public 52.23 52.645 34

28 low private 52.23 52.645 34

29 middle private 52.23 52.645 34

30 high private 52.23 52.645 34

31 low public 52.23 52.645 35

32 middle public 52.23 52.645 35

33 high public 52.23 52.645 35

34 low private 52.23 52.645 35

35 middle private 52.23 52.645 35

36 high private 52.23 52.645 35

37 low public 52.23 52.645 36

38 middle public 52.23 52.645 36

39 high public 52.23 52.645 36

40 low private 52.23 52.645 36

41 middle private 52.23 52.645 36

42 high private 52.23 52.645 36

43 low public 52.23 52.645 37

44 middle public 52.23 52.645 37

45 high public 52.23 52.645 37

46 low private 52.23 52.645 37

47 middle private 52.23 52.645 37

48 high private 52.23 52.645 37

49 low public 52.23 52.645 38

50 middle public 52.23 52.645 38

51 high public 52.23 52.645 38

52 low private 52.23 52.645 38

53 middle private 52.23 52.645 38

54 high private 52.23 52.645 38

55 low public 52.23 52.645 39

56 middle public 52.23 52.645 39

57 high public 52.23 52.645 39

58 low private 52.23 52.645 39

59 middle private 52.23 52.645 39

60 high private 52.23 52.645 39

61 low public 52.23 52.645 40

62 middle public 52.23 52.645 40

63 high public 52.23 52.645 40

64 low private 52.23 52.645 40

65 middle private 52.23 52.645 40

66 high private 52.23 52.645 40

67 low public 52.23 52.645 41

68 middle public 52.23 52.645 41

69 high public 52.23 52.645 41

70 low private 52.23 52.645 41

71 middle private 52.23 52.645 41

72 high private 52.23 52.645 41

73 low public 52.23 52.645 42

74 middle public 52.23 52.645 42

75 high public 52.23 52.645 42

76 low private 52.23 52.645 42

77 middle private 52.23 52.645 42

78 high private 52.23 52.645 42

79 low public 52.23 52.645 43

80 middle public 52.23 52.645 43

81 high public 52.23 52.645 43

82 low private 52.23 52.645 43

83 middle private 52.23 52.645 43

84 high private 52.23 52.645 43

85 low public 52.23 52.645 44

86 middle public 52.23 52.645 44

87 high public 52.23 52.645 44

88 low private 52.23 52.645 44

89 middle private 52.23 52.645 44

90 high private 52.23 52.645 44

91 low public 52.23 52.645 45

92 middle public 52.23 52.645 45

93 high public 52.23 52.645 45

94 low private 52.23 52.645 45

95 middle private 52.23 52.645 45

96 high private 52.23 52.645 45

97 low public 52.23 52.645 46

98 middle public 52.23 52.645 46

99 high public 52.23 52.645 46

100 low private 52.23 52.645 46

101 middle private 52.23 52.645 46

102 high private 52.23 52.645 46

103 low public 52.23 52.645 47

104 middle public 52.23 52.645 47

105 high public 52.23 52.645 47

106 low private 52.23 52.645 47

107 middle private 52.23 52.645 47

108 high private 52.23 52.645 47

109 low public 52.23 52.645 48

110 middle public 52.23 52.645 48

111 high public 52.23 52.645 48

112 low private 52.23 52.645 48

113 middle private 52.23 52.645 48

114 high private 52.23 52.645 48

115 low public 52.23 52.645 49

116 middle public 52.23 52.645 49

117 high public 52.23 52.645 49

118 low private 52.23 52.645 49

119 middle private 52.23 52.645 49

120 high private 52.23 52.645 49

121 low public 52.23 52.645 50

122 middle public 52.23 52.645 50

123 high public 52.23 52.645 50

124 low private 52.23 52.645 50

125 middle private 52.23 52.645 50

126 high private 52.23 52.645 50

127 low public 52.23 52.645 51

128 middle public 52.23 52.645 51

129 high public 52.23 52.645 51

130 low private 52.23 52.645 51

131 middle private 52.23 52.645 51

132 high private 52.23 52.645 51

133 low public 52.23 52.645 52

134 middle public 52.23 52.645 52

135 high public 52.23 52.645 52

136 low private 52.23 52.645 52

137 middle private 52.23 52.645 52

138 high private 52.23 52.645 52

139 low public 52.23 52.645 53

140 middle public 52.23 52.645 53

141 high public 52.23 52.645 53

142 low private 52.23 52.645 53

143 middle private 52.23 52.645 53

144 high private 52.23 52.645 53

145 low public 52.23 52.645 54

146 middle public 52.23 52.645 54

147 high public 52.23 52.645 54

148 low private 52.23 52.645 54

149 middle private 52.23 52.645 54

150 high private 52.23 52.645 54

151 low public 52.23 52.645 55

152 middle public 52.23 52.645 55

153 high public 52.23 52.645 55

154 low private 52.23 52.645 55

155 middle private 52.23 52.645 55

156 high private 52.23 52.645 55

157 low public 52.23 52.645 56

158 middle public 52.23 52.645 56

159 high public 52.23 52.645 56

160 low private 52.23 52.645 56

161 middle private 52.23 52.645 56

162 high private 52.23 52.645 56

163 low public 52.23 52.645 57

164 middle public 52.23 52.645 57

165 high public 52.23 52.645 57

166 low private 52.23 52.645 57

167 middle private 52.23 52.645 57

168 high private 52.23 52.645 57

169 low public 52.23 52.645 58

170 middle public 52.23 52.645 58

171 high public 52.23 52.645 58

172 low private 52.23 52.645 58

173 middle private 52.23 52.645 58

174 high private 52.23 52.645 58

175 low public 52.23 52.645 59

176 middle public 52.23 52.645 59

177 high public 52.23 52.645 59

178 low private 52.23 52.645 59

179 middle private 52.23 52.645 59

180 high private 52.23 52.645 59

181 low public 52.23 52.645 60

182 middle public 52.23 52.645 60

183 high public 52.23 52.645 60

184 low private 52.23 52.645 60

185 middle private 52.23 52.645 60

186 high private 52.23 52.645 60

187 low public 52.23 52.645 61

188 middle public 52.23 52.645 61

189 high public 52.23 52.645 61

190 low private 52.23 52.645 61

191 middle private 52.23 52.645 61

192 high private 52.23 52.645 61

193 low public 52.23 52.645 62

194 middle public 52.23 52.645 62

195 high public 52.23 52.645 62

196 low private 52.23 52.645 62

197 middle private 52.23 52.645 62

198 high private 52.23 52.645 62

199 low public 52.23 52.645 63

200 middle public 52.23 52.645 63

201 high public 52.23 52.645 63

202 low private 52.23 52.645 63

203 middle private 52.23 52.645 63

204 high private 52.23 52.645 63

205 low public 52.23 52.645 64

206 middle public 52.23 52.645 64

207 high public 52.23 52.645 64

208 low private 52.23 52.645 64

209 middle private 52.23 52.645 64

210 high private 52.23 52.645 64

211 low public 52.23 52.645 65

212 middle public 52.23 52.645 65

213 high public 52.23 52.645 65

214 low private 52.23 52.645 65

215 middle private 52.23 52.645 65

216 high private 52.23 52.645 65

217 low public 52.23 52.645 66

218 middle public 52.23 52.645 66

219 high public 52.23 52.645 66

220 low private 52.23 52.645 66

221 middle private 52.23 52.645 66

222 high private 52.23 52.645 66

223 low public 52.23 52.645 67

224 middle public 52.23 52.645 67

225 high public 52.23 52.645 67

226 low private 52.23 52.645 67

227 middle private 52.23 52.645 67

228 high private 52.23 52.645 67

229 low public 52.23 52.645 68

230 middle public 52.23 52.645 68

231 high public 52.23 52.645 68

232 low private 52.23 52.645 68

233 middle private 52.23 52.645 68

234 high private 52.23 52.645 68

235 low public 52.23 52.645 69

236 middle public 52.23 52.645 69

237 high public 52.23 52.645 69

238 low private 52.23 52.645 69

239 middle private 52.23 52.645 69

240 high private 52.23 52.645 69

241 low public 52.23 52.645 70

242 middle public 52.23 52.645 70

243 high public 52.23 52.645 70

244 low private 52.23 52.645 70

245 middle private 52.23 52.645 70

246 high private 52.23 52.645 70

247 low public 52.23 52.645 71

248 middle public 52.23 52.645 71

249 high public 52.23 52.645 71

250 low private 52.23 52.645 71

251 middle private 52.23 52.645 71

252 high private 52.23 52.645 71

253 low public 52.23 52.645 72

254 middle public 52.23 52.645 72

255 high public 52.23 52.645 72

256 low private 52.23 52.645 72

257 middle private 52.23 52.645 72

258 high private 52.23 52.645 72

259 low public 52.23 52.645 73

260 middle public 52.23 52.645 73

261 high public 52.23 52.645 73

262 low private 52.23 52.645 73

263 middle private 52.23 52.645 73

264 high private 52.23 52.645 73

265 low public 52.23 52.645 74

266 middle public 52.23 52.645 74

267 high public 52.23 52.645 74

268 low private 52.23 52.645 74

269 middle private 52.23 52.645 74

270 high private 52.23 52.645 74

271 low public 52.23 52.645 75

272 middle public 52.23 52.645 75

273 high public 52.23 52.645 75

274 low private 52.23 52.645 75

275 middle private 52.23 52.645 75

276 high private 52.23 52.645 75

277 low public 52.23 52.645 76

278 middle public 52.23 52.645 76

279 high public 52.23 52.645 76

280 low private 52.23 52.645 76

281 middle private 52.23 52.645 76

282 high private 52.23 52.645 76

283 low public 52.23 52.645 77

284 middle public 52.23 52.645 77

285 high public 52.23 52.645 77

286 low private 52.23 52.645 77

287 middle private 52.23 52.645 77

288 high private 52.23 52.645 77

289 low public 52.23 52.645 78

290 middle public 52.23 52.645 78

291 high public 52.23 52.645 78

292 low private 52.23 52.645 78

293 middle private 52.23 52.645 78

294 high private 52.23 52.645 78

295 low public 52.23 52.645 79

296 middle public 52.23 52.645 79

297 high public 52.23 52.645 79

298 low private 52.23 52.645 79

299 middle private 52.23 52.645 79

300 high private 52.23 52.645 79

301 low public 52.23 52.645 80

302 middle public 52.23 52.645 80

303 high public 52.23 52.645 80

304 low private 52.23 52.645 80

305 middle private 52.23 52.645 80

306 high private 52.23 52.645 80

> ex8preds <- predict(ex6, ex8data, "probs")

> sapply(levels(student$ses), function(s){apply(data.frame(ex8preds[ex8data$ses == s, ]), 2, mean)})

low middle high

general 0.4067122 0.3096582 0.24457624

academic 0.5124482 0.5063960 0.66794338

vocation 0.0808396 0.1839458 0.08748038

>