Discrete Multivariate Analysis - 579 - Final Exam - Part 2

Alex Towell (atowell@siue.edu)

Applicants for graduate school are classified according to department, sex, and admission status. A goal of the study is to determine the role an applicant's sex plays in the determination of admission status.

```
# department and sex are defined through indicator variables
grad.data <- data.frame(dep=c(0,0,1,1),
                        sex=c(0,1,0,1),
                        yes=c(235,38,122,103),
                        no=c(35,7,93,69))

# define row sample sizes
grad.data$n = grad.data$yes + grad.data$no
# define the interaction variable
grad.data$dep.sex = grad.data$dep * grad.data$sex

print(grad.data)
```

```
##   dep sex yes no   n dep.sex
## 1   0   0 235 35 270       0
## 2   0   1  38  7  45       0
## 3   1   0 122 93 215       0
## 4   1   1 103 69 172       1
```

Problem 1

Define a main effects logistic regression model M having two binary input variables. Include notation for the design matrix X, and the parameter vector β. Provide an interpretation for each of the effect parameters in β, stated in the context of the problem.

The data is cross-sectional data given by

$$(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x_n}, y_n)$$

where $\boldsymbol{x_i} = (1, x_{i1}, x_{i2})'$.

The response variables $\{y_1, \dots, y_n\}$ are given by

$$y = \begin{cases} 1 & \text{if admit=yes} \\ 0 & \text{if admit=no} \end{cases}$$

The explanatory indicator variables x_1 and x_2 are respectively given by

$$x_1 = \begin{cases} 1 & \text{if department=non-science} \\ 0 & \text{if department=science} \end{cases}$$

and

$$x_2 = \begin{cases} 1 & \text{if sex=female} \\ 0 & \text{if sex=male} \end{cases}$$

Then, π denotes the probability of being admitted,

$$\pi(\boldsymbol{x}) = \Pr(y = \text{yes} \mid \boldsymbol{x}).$$

That is, π models probability as a function of $\boldsymbol{x}$.

The probability model is given by

$$y_i \sim \text{BIN}(1, \pi(\boldsymbol{x_i}))$$

The main effects model M is given by

$$\text{logit}(\pi(\boldsymbol{x})) = \boldsymbol{x}'\boldsymbol{\beta},$$

where the parameter vector $\boldsymbol{\beta}$ of dimension 3×1 is given by

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

.

The parameter β_1 is the partial effect of department (non-science - science) on admission and the parameter β_2 is the partial effect of sex (female - male) on admission.

In the main effects model M, we assume the effect of x_1 is the same at both levels of x_2 and vice versa, i.e.,

$$\theta(\text{science} : \text{non-science} \mid \text{male}) = \theta(\text{science} : \text{non-science} \mid \text{female}) = \exp(\beta_1)$$

and

$$\theta(\text{male} : \text{female} \mid \text{science}) = \theta(\text{male} : \text{female} \mid \text{non-science}) = \exp(\beta_2).$$

The logistic regression model for M is given by

$$\text{logit}(\pi(\boldsymbol{x})) = \boldsymbol{x_i'}\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2},$$

such that if we solve for $\pi(\boldsymbol{x_i})$ we get the result

$$\pi(\boldsymbol{x_i}) = \frac{\exp(\boldsymbol{x_i'}\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x_i'}\boldsymbol{\beta})}.$$

The design matrix is given by

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}.$$

Since we are dealing with Boolean input variables $\boldsymbol{x} = (1, x_1, x_2)$, we only need to count the number of times each discrete input in the design matrix occurs, say n_j for the j-th row, denoted by $\boldsymbol{x_j}$, and let the corresponding y_j denote the number of times the response was *yes* for $\boldsymbol{x_j}$.

Then, the probabilistic model for y_j is the binomial distribution,

$$y_j \sim \text{BIN}(n_j, \pi(\boldsymbol{x_j}))$$

and we estimate $\boldsymbol{\beta}$ by maximizing

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \prod_{i=1}^{4} \pi_i(\boldsymbol{\beta})^{y_i}(1-\pi_i(\boldsymbol{\beta}))^{n_i - y_i}.$$

Problem 2

Provide notation for the design matrix X_S and parameter vector β_S for the saturated model M_S. Provide a brief description of an interaction effect.

For the saturated model M_S (interaction model), we must include all possible effects and interactions, which means relative to the main effects model M we add account for interaction effects between x_1 and x_2, i.e.,

$$\text{logit}(\pi(\boldsymbol{x})) = \boldsymbol{x}'\boldsymbol{\beta_S} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

where the parameter vector $\boldsymbol{\beta_S}$ of dimension 4×1 is given by

$$\boldsymbol{\beta_S} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

.

The interaction effect between department and sex (x_1 and x_2) occurs when the effect of an input depends on the level of the input of the other input. In the main effects model M, we assume the effect of x_1 is the same at both levels of x_2 and vice versa, i.e.,

$$\begin{aligned} \theta(\text{science} : \text{non-science} \mid \text{male}) &= e^{\beta_1}, \\ \theta(\text{science} : \text{non-science} \mid \text{female}) &= e^{\beta_1 + \beta_3}, \\ \theta(\text{male} : \text{female} \mid \text{science}) &= e^{\beta_2}, \\ \theta(\text{male} : \text{female} \mid \text{non-science}) &= e^{\beta_2 + \beta_3}. \end{aligned}$$

The design matrix for M_S is given by

$$\mathbf{X_S} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

Problem 3

For each of the models M_O, M_1, M_2, provide notation for the design matrix and a brief description of the model effects, stated in the context of the problem.

The design matrix for M_O, the independence model, is given by

$$\mathbf{X_O} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

This model assumes that neither department nor sex has an effect on admission, i.e., $\text{logit}(\pi) = \beta_0$.

The design matrix for M_1 is given by

$$\mathbf{X_1} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

This model considers only the marginal effect of department on admission, i.e., $\text{logit}(\pi(x_1)) = \beta_0 + \beta_1 x_1$.

The design matrix for M_2 is given by

$$\mathbf{X_2} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}.$$

This model considers only the marginal effect of sex on admission, i.e., $\text{logit}(\pi(x_1)) = \beta_0 + \beta_2 x_2$.

Problem 4

Compute the deviance statistic D, and give degrees of freedom Δdf, for each of the models M_O, M_1, M_2, M, M_S from the grad school data. Provide a general form for the statistic G^2, and the degrees of freedom for the reference chi-square distribution, for testing a reduced model M_R against a full model M_F.

$G^2(M \mid M_S) = D(M)$ is called the deviance for testing model M goodness-of-fit.

The likelihood ratio statistic G^2 for testing a reduced model M_R against a full model M_F is given by

$$\begin{aligned} G^2(M_R \mid M_S) &= [-2L_R] - [-2L_F] \\ &= ([-2L_R] - [-2L_S]) - ([-2L_F] - [-2L_S]) \\ &= G^2(M_R \mid M_S) - G^2(M_F \mid M_S) \\ &= D(M_R) - D(M_F). \end{aligned}$$

The degree of freedom is given by

$$df = P_F - P_R = (P_S - P_R) - (P_S - P_F) = \Delta df_R - \Delta df_F,$$

and so the reference distribution is χ^2 with $df = \Delta df_R - \Delta df_F$.

```
# fit each of the candidate models
m.s = glm(yes/n ~ dep+sex+sex*dep,weights=n,family=binomial,data=grad.data)
m.12 = glm(yes/n ~ dep+sex,weights=n,family=binomial,data=grad.data)
m.1 = glm(yes/n ~ dep,weights=n,family=binomial,data=grad.data)
m.2 = glm(yes/n ~ sex,weights=n,family=binomial,data=grad.data)
m.0 = glm(yes/n ~ 1,weights=n,family=binomial,data=grad.data)

# create a table for candidate model deviances
# the rows represent the model, the deviance, and the degrees of freedom
deviance.table=matrix(c(
  0,m.12$deviance,m.1$deviance,m.2$deviance,m.0$deviance,
  0,m.12$df.residual,m.1$df.residual,m.2$df.residual,m.0$df.residual),
  nrow=5)
dimnames(deviance.table) = list(c("MS","M","M1","M2","MO"),c("deviance","delta.df"))
print(deviance.table)
```

```
##      deviance delta.df
## MS  0.0000000        0
## M   0.4589638        1
## M1  0.6036551        2
## M2 67.8794451        2
## MO 73.1962870        3
```

Problem 5

Compute the likelihood statistic G^2 for testing reduced model M against full model M_S from the grad school data, and provide an interpretation in the context of the problem.

```
# test the main effects model against the interaction (saturated) model
anova(m.12,m.s,test="LRT")
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	0.4589638	NA	NA	NA
0	0.0000000	1	0.4589638	0.4981086

The statistic $G^2(M \mid M_S) = D(M) = 0.459$, which has a p-value of 0.498. Thus, we conlcude the main effects model (no interaction on inputs) is compatible with the observed data.

Problem 6

Compute the likelihood statistic G^2 for testing reduced model M_O against full model M_2 from the grad school data, and provide an interpretation in the context of the problem.

We are testing for a marginal effect of x_2 (sex), ignoring the effects of x_1 (department).

```
# test for the input 2 effect using a marginal effect test
anova(m.0,m.2,test="LRT")
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
3	73.19629	NA	NA	NA
2	67.87945	1	5.316842	0.0211203

The statistic $G^2(M_0 \mid M_2) = 5.317$ with $df = 1$ has a p-value around 0.021. Thus, we conlcude that the data supports the model where sex has a marginal effect on admission.

Problem 7

Compute the likelihood statistic G^2 for testing reduced model M_1 against full model M from the grad school data, and provide an interpretation in the context of the problem. Include an explanation of how this test differs from that of the previous problem.

We are testing for a partial effect of x_2 (sex), adjusting for the effect of x_1 (department).

```
anova(m.1,m.12,test="LRT")
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
2	0.6036551	NA	NA	NA
1	0.4589638	1	0.1446913	0.7036611

The statistic $G^2(M_1 \mid M) = 67.420$ with $df = 1$ has p-value $= 0.704$, thus we conclude that when we adjust for x_1 (department), the data supports the model where x_2 (sex) does not have a significant effect.

How are we to interpret the results? We have shown that sex has a marginal effect on admission, but its effect when adjusting for department is neglibible.

We make the following observation. Males are more likely to choose science, and science department is more likely to accept. Females are more likely to choose non-science, and non-science department is less likely to accept. Thus, we see that overall, females are less likely to be admitted, thus showing that sex has a marginal effect on admission. However, the probability of admission is the same for both sexes when we adjust for department.

Thus, model M_1 (department) is the major effect.

Problem 8

Compute estimates of the response probabilities based on model M_1 from the grad school data, and provide an interpretation in the context of the problem.

```
# compute probability estimates under model M1
pred.1 = predict(m.1,type = "response")
```

```
prob.table = matrix(c(pred.1),nrow = 4)
dimnames(prob.table) = list(c("science male","science female",
                              "non-science male","non-science female"),
                            c("prob.1"))
print(prob.table,digits = 4)
```

```
##                    prob.1
## science male       0.8667
## science female     0.8667
## non-science male   0.5814
## non-science female 0.5814
```

Under model M_1, we estimate the probability of admission into the science department is 0.8667 and we estimate probability of admission into the non-science department is 0.5814.

Note that under this model, sex has no effect on the probability of admission. This model is justified on the basis of the data and the previous tests we conducted.