Discrete Multivariate Analysis - 579 - Final Exam

Alex Towell (atowell@siue.edu)

Contents

1 Part 1

A sample of elderly patients were given a psychiatric examination to determine whether symptoms of senility are present. One explanatory variable is a patient's score on the Wechsler Adult Intelligence Scale (WAIS).

```
wais.data <- data.frame(
  wais=c(4,5,6,7,8,9,10,11,12,13,14,15,16,17,18),
  n=c(2,1,2,3,2,6,6,6,2,6,7,3,4,1,1),
  senile=c(1,1,1,2,2,2,2,1,0,1,2,0,0,0,0))

print(wais.data)
```

```
##    wais n senile
## 1     4 2      1
## 2     5 1      1
## 3     6 2      1
## 4     7 3      2
## 5     8 2      2
```

```
## 6      9 6      2
## 7     10 6      2
## 8     11 6      1
## 9     12 2      0
## 10    13 6      1
## 11    14 7      2
## 12    15 3      0
## 13    16 4      0
## 14    17 1      0
## 15    18 1      0
```

1.1 Problem 1

Define the logistic regression model, including notation for the input matrix X, the response vector y, the parameter vector β, and the probability vector $\pi(\beta)$.

The data is $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x_n}, y_n)$ where $\boldsymbol{x_j}$ is a column vector of explanatory variables and y_j is a binary variable.

The probability model is given by

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

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

The logistic regression model is given by

$$\log\left(\frac{\pi(\boldsymbol{x_i})}{1-\pi(\boldsymbol{x_i})}\right) = \boldsymbol{x_i'}\boldsymbol{\beta},$$

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

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

where $\boldsymbol{x_i'}$ denotes the transpose of $\boldsymbol{x_i}$.

The response vector $\boldsymbol{y}$ of dimension $n \times 1$ is given by

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

The input (design) martix $\mathbf{X}$ of dimension $n \times (p+1)$ is given by

$$\mathbf{X} = \begin{pmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x_n'} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}.$$

The parameter vector $\boldsymbol{\beta}$ of dimension $(p+1) \times 1$ is given by

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

where β_0 denotes the *intercept* in the linear predictor.

The probability vector

$$\boldsymbol{\pi}(\boldsymbol{\beta}) = \begin{pmatrix} \pi(\boldsymbol{x}_1) \\ \pi(\boldsymbol{x}_2) \\ \vdots \\ \pi(\boldsymbol{x_n}) \end{pmatrix}$$

where $\boldsymbol{x_j}$ are the explanatory varibles as described previously.

1.2 Problem 2

1.2.1 Part (a)

State the likelihood equation for the MLE $\hat{\boldsymbol{\beta}}$ as a normal equation.

The normal equations for $\hat{\boldsymbol{\beta}}$ are given by

$$\mathbf{X}'(\boldsymbol{y} - \boldsymbol{\pi}(\hat{\boldsymbol{\beta}})) = 0.$$

1.2.2 Part (b)

State the equation for $\hat{V} = \hat{\text{Cov}}(\hat{\beta})$, the estimated variance matrix for $\hat{\beta}$.

The equation for $\hat{\mathbf{V}}$ is given by

$$\hat{\mathbf{V}} = \mathbf{c\hat{o}v}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1},$$

which is a $(p+1) \times (p+1)$ covariance matrix (estimate) and $\hat{\mathbf{W}} = \mathbf{diag}(\pi_i(1-\pi_i))$ is an $n \times n$ diagonal matrix.

1.2.3 Part (c)

Compute $\hat{\beta}$ and $\hat{V}$ from the WAIS data.

```r
wais.mod = glm(senile/n ~ wais,weights=n, family=binomial, data=wais.data)

# define the intercept and slope parameter estimates
# alpha
a = wais.mod$coefficients[1]
# beta
b = wais.mod$coefficients[2]

# this is beta = (a,b)'
print(c(a,b))
```

```
## (Intercept)        wais
##   2.4810957  -0.3189723
```

```
# compute the variance/covariance matrix for parameter estimates
V.hat = vcov(wais.mod)
print(V.hat)
```

```
##             (Intercept)        wais
## (Intercept)   1.4447178 -0.13154006
## wais         -0.1315401  0.01301779
```

We see that $\hat{\boldsymbol{\beta}} = (2.481, -0.319)'$ and

$$\hat{\mathbf{V}} = \begin{pmatrix} 1.445 & -0.132 \\ -0.132 & 0.013 \end{pmatrix}.$$

1.3 Problem 3

1.3.1 Part (a)

State the equation for a 95% confidence interval for β_j.

A 95% confidence interval for β_j is $\hat{\beta}_j \pm 1.96\sqrt{\hat{V}_{jj}}$, where $\hat{V}_{jj}$ is the j-th element along the diagonal of $\hat{\mathbf{V}}$.

1.3.2 Part (b)

Compute a 95% confidence interval for β_1 from the WAIS data, and provide an interpretation in the context of the problem.

```
# compute the s.e. for beta.hat from the estimated variance/covariance matrix
se.b = sqrt(V.hat[2,2])

# compute a 95% confidence interval estimate for beta
L.beta = b - 1.96*se.b
U.beta = b + 1.96*se.b
print(b)
```

```
##       wais
## -0.3189723
```

```
print(c(L.beta,U.beta))
```

```
##       wais       wais
## -0.5425995 -0.0953451
```

We estimate that β_1 is between -0.543 and -0.095.

Note that β_1 can generally be thought of as an effect size for the association between WAIS and senility, in particular the change in the log odds with respect to a a change in the input level of WAIS.

We estimate that the log-odds of senility *decreases* by 0.319 units from a unit *increase* in the WAIS measure, which makes sense.

1.4 Problem 4

1.4.1 Part (a)

State the equations for a 95% confidence interval for odds ratio θ.

The *odds* at $\boldsymbol{x}$ are given by

$$\Omega(\boldsymbol{x}) = \frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})} = \exp(\boldsymbol{x}'\boldsymbol{\beta}).$$

Then,

$$\log \Omega(\boldsymbol{x}) = \log\left(\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}\right) = \boldsymbol{x}'\boldsymbol{\beta}.$$

Let $\boldsymbol{u_j}$ be defined as a unit vector of dimension $p+1$ such that every element except the j-th element is zero.

Then

$$\begin{aligned} \beta_j &= \log \Omega(\boldsymbol{x} + \boldsymbol{u_j}) - \log \Omega(\boldsymbol{x}) \\ &= \log \frac{\Omega(\boldsymbol{x} + \boldsymbol{u_j})}{\Omega(\boldsymbol{x})}, \end{aligned}$$

which is the log odds ratio for comparing inputs $\boldsymbol{x}$ and $\boldsymbol{x} + \boldsymbol{u_j}$. So, $\exp(\beta_j)$ is the odds ratio θ_j for comparing inputs $\boldsymbol{x}$ and $\boldsymbol{x} + \boldsymbol{u_j}$.

1.4.2 Simple logistic regression

Since $p = 2$, we only have one explanatory variable x and thus we may rephrase this as β_1 is the log odds ratio for comparing input levels x and $x+1$ and therefore $\exp(\beta_1)$ is the odds ratio θ for comparing input levels x and $x+1$.

A 95% confidence interval for θ is given by

$$\left[\exp(l_{\beta_1}), \exp(u_{\beta_1})\right]$$

where $l_{\beta_1} = \hat{\beta}_1 - 1.96\sqrt{\hat{V}_{11}}$ and $u_{\beta_1} = \hat{\beta}_1 + 1.96\sqrt{\hat{V}_{11}}$.

Note that we use zero-based indexing, i.e., the first element in the matrix is at index $(0,0)$ instead of $(1,1)$.

1.4.3 Part (b)

Compute a 95% confidence interval for θ from the WAIS data.

```r
# compute an estimate of the odds ratio
theta <- exp(b)

# compute a 95% confidence interval estimate for the odds ratio
```

```
L.theta = exp(L.beta)
U.theta = exp(U.beta)

# print confidence interval
print(c(L.theta,U.theta))
```

```
##      wais      wais
## 0.5812354 0.9090592
```

Since the log-odds is given by β_1, the odds θ is given by $\exp(\beta_1)$ with a point estimate $\hat{\theta} = \exp(\hat{\beta}_1)$ given by

$$0.727.$$

From the R computation, we see that a 95% confidence interval for θ is given by

$$(0.581, 0.909),$$

which is not symmetric around the point estimate.

1.5 Problem 5

1.5.1 Part (a)

State the equations for a 95% confidence interval for the logit L_o at input level x_o:

We estimate the logit with

$$\hat{L}_o = \boldsymbol{x}_o' \hat{\boldsymbol{\beta}}$$

which has a variance

$$\operatorname{Var}(\hat{L}_o) = \boldsymbol{x}_o' \mathbf{V} \boldsymbol{x}_o$$

where $\mathbf{V} = \operatorname{Cov}(\hat{\boldsymbol{\beta}})$.

We do not know $\mathbf{V}$, so we estimate $\sigma_{\hat{L}_o}$ with

$$\hat{\sigma}(\hat{L}_o) = \sqrt{\boldsymbol{x}_o' \hat{\mathbf{V}} \boldsymbol{x}_o}.$$

Thus, a 95% confidence interval for L_o is given by

$$\hat{L}_o \pm 1.96 \hat{\sigma}(\hat{L}_o).$$

1.5.1.1 Simple logistic regression

Since this is simple logistic regression with $p = 1$ explanatory variables, we may simplify the presentation.

Let $\boldsymbol{x}_o' = (1, x_o)$. Then, these equations simplify to

$$\hat{L}_o = \hat{\beta}_0 + \hat{\beta}_1 x_o,$$

and

$$\operatorname{Var}(\hat{L}_o) = \operatorname{Var}(\hat{\beta}_0) + x_o^2 \operatorname{Var}(\hat{\beta}_1) + 2 x_0 \operatorname{Cov}(\hat{\beta}_1, \hat{\beta}_1).$$

$$\sigma^2_{\hat{L}_o} = \hat{V}_{00} + x_o^2\hat{V}_{11} + 2x_o\hat{V}_{01}.$$

$$\begin{aligned}\hat{\sigma}(\hat{L}_o) &= \sqrt{\mathrm{Var}(\hat{\beta}_0) + x_o^2\,\mathrm{Var}() + 2x_o\hat{V}_{01}} \\ &= \sqrt{\hat{V}_{00} + x_o^2\hat{V}_{11} + 2x_o\hat{V}_{01}}.\end{aligned}$$

A 95% CI for L_o is given by

$$(1, x_o)^t\beta \pm 1.96\sqrt{(1, x_o)\hat{V}(1, x_o)^t}.$$

1.5.2 Part (b)

Compute a 95% confidence interval for L_o at input level $x_o = 10$ from the WAIS data.

```
xo <- 10
L.hat <- a + b*xo # should agree with L0.hat below

x0 = as.matrix(c(1,xo))
print(x0)
```

```
##      [,1]
## [1,]    1
## [2,]   10
```

```
beta.hat = as.matrix(c(a,b))

L0.hat = t(x0) %*% beta.hat
se0 = sqrt(t(x0)%*%V.hat%*%x0)
print(c(L0.hat,se0))
```

```
## [1] -0.7086274  0.3401400
```

```
L.L0 = L0.hat - 1.96*se0
U.L0 = L0.hat + 1.96*se0
print(c(L.L0,U.L0))
```

```
## [1] -1.37530182 -0.04195301
```

1.6 Problem 6

1.6.1 Part (a)

State the equations for a 95% confidence interval for the probability π_o at input level x_o.

1.6.2 Part (b)

Compute a 95% confidence interval for π_o at input level $x_o = 10$ from the WAIS data, and provide an interpretation in the context of the problem.

```
pi.hat <- exp(L.hat) / (1+exp(L.hat))
#print(pi.hat)
print(c(exp(L.LO) / (1+exp(L.LO)),exp(U.LO) / (1+exp(U.LO))))
```

```
## [1] 0.2017646 0.4895133
```

2 Part 2

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.

2.1 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.

$\text{logit}(\pi)(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & & \\ 1 & x_{21} & x_{22} & & \\ \vdots & \vdots & \vdots 1 & x_{n1} & x_{n2} \end{pmatrix}$$

```
# 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
```

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



#compute probability estimates under each of the candidate models
pred.S = predict(m.s,type = "response")
pred.12 = predict(m.12,type = "response")
pred.1 = predict(m.1,type = "response")
pred.2 = predict(m.2,type = "response")
pred.0 = predict(m.0,type = "response")
```

2.2 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.

2.3 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.

2.4 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.

```r
# 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 67.8794451        2
## M2  0.6036551        2
## MO 73.1962870        3
```

First, $G^2(M \,|\, M_S) = D(M)$ is called the deviance of M.

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

$$\begin{aligned} G^2(M_R \,|\, M_S) &= [-2L_R] - [-2L_F] && (1) \\ &= ([-2L_R] - [-2L_S]) - ([-2L_F] - [-2L_S]) && (2) \\ &= G^2(M_R \,|\, M_S) - G^2(M_F \,|\, M_S) && (3) \\ &= D(M_R) - D(M_F). && (4) \end{aligned}$$

The df 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 if χ^2 with $df = \Delta df_R - \Delta df_F$ degrees of freedom.

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

2.6 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.

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

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
3	73.1962870	NA	NA	NA
2	0.6036551	1	72.59263	0

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

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
2	67.8794451	NA	NA	NA
1	0.4589638	1	67.42048	0

2.7 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.

```
#test the independence model against the main effects model
anova(m.0,m.12,test = "LRT")
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
3	73.1962870	NA	NA	NA
1	0.4589638	2	72.73732	0

```
#test for the input 1 effect, first using a marginal effect test, then using a partial effect
anova(m.0,m.1,test = "LRT")
```

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

```
anova(m.2,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

2.8 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.

```
#create a display a table for the probability estimates
prob.table = matrix(c(pred.S,pred.12,pred.1,pred.2,pred.0),nrow = 4)
#dimnames(prob.table) = list(c("female low","female high","male low","male high"),
#                            c("prob.S","prob.12","prob.1","prob.2","prob.0"))

print(prob.table,digits = 4)
```

```
##        [,1]   [,2]   [,3]   [,4]   [,5]
## [1,] 0.8704 0.8655 0.7361 0.8667 0.7094
## [2,] 0.8444 0.8737 0.6498 0.8667 0.7094
## [3,] 0.5674 0.5736 0.7361 0.5814 0.7094
## [4,] 0.5988 0.5912 0.6498 0.5814 0.7094
```

```
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
```