Discrete Multivariate Analysis - 579 - Final Exam - Part 1

Alex Towell (atowell@siue.edu)

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 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 response 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

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

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

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

where $\boldsymbol{x}_{\boldsymbol{i}}'$ 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

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

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 variables as described previously and the i-th element of $\boldsymbol{\pi}(\boldsymbol{\beta})$ may be denoted by $\pi_i(\boldsymbol{\beta})$.

2 Problem 2

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

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

2.2 Part (b)

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

The $(p+1) \times (p+1)$ covariance matrix $\text{Cov}(\hat{\boldsymbol{\beta}})$ has an estimator $\hat{\mathbf{V}}$ given by

$$\hat{\mathbf{V}} = (\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1},$$

where

$$\hat{\mathbf{W}} = \mathbf{diag}\{\pi_i(\hat{\boldsymbol{\beta}})(1 - \pi_i(\hat{\boldsymbol{\beta}}))\}$$

is an $n \times n$ diagonal weight matrix.

2.3 Part (c)

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

```
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

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

3 Problem 3

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 $\hat{\mathbf{V}}$.

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.

First, we justify a particular interpretation of β_j in the logistic regression model. 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 $\boldsymbol{x}+\boldsymbol{u_j}$.

Thus, we see that β_j can be thought of as an effect size for the association between x_j and y, in particular the change in the log odds ratio with respect to a unit increase in the input level of x_j.

Since $p=1$ we only have one explanatory variable `WAIS` and thus we may rephrase this as β_1 is the log odds ratio for comparing `WAIS` at input levels x and $x+1$.

Another way to say this is β_1 is the change in the difference of the log odds for comparing `WAIS` at input levels x and $x+1$.

We compute the confidence interval using the following R code.

```
# 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. We estimate that the log-odds ratio of senility *decreases* by at least 0.095 and at worst 0.543 given a unit *increase* in the WAIS measure.

4 Problem 4

4.1 Part (a)

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

Since β_1 is the log-odds ratio when comparing x to $x+1$ $\theta = \exp(\beta_1)$ is the odds ratio for comparing x to $x+1$.

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 $(1,1)$.

4.2 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
```

We see that a 95% confidence interval for θ is given by

$$(0.581, 0.909).$$

5 Problem 5

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

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

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

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

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

Since this is simple logistic regression with $p = 1$ explanatory variables, we let $\boldsymbol{x}_o' = (1, x_o)$ which yields the simplifications

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

and

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

5.2 Part (b)

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

```
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
```

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

$$(-1.375, -0.042).$$

6 Problem 6

6.1 Part (a)

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

Let L_o be the logit at $\boldsymbol{x} = \boldsymbol{x_o}$, i.e.,

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

We would like to estimate $\pi(\boldsymbol{x_o})$. We let π_o denote $\pi(\boldsymbol{x_o})$, thus

$$\pi_o = \frac{\exp(L_o)}{1+\exp(L_o)}.$$

A 95% confidence interval for π_o is given by

$$\left[\frac{\exp(\ell_{L_o})}{1+\exp(\ell_{L_o})}, \frac{\exp(u_{L_o})}{1+\exp(u_{L_o})}\right]$$

where ℓ_{L_o} and u_{L_o} are respectively the lower and upper bounds of the confidence interval for L_o.

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.

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

```
## 0.2017646 0.4895133
```

Thus, we see that a 95% confidence interval for π_o is given by

$$[0.2018, 0.4895].$$

We estimate that the probability of senility given a WAIS score of 10 is between 0.2018 and 0.4895.