Homework 4

Louis Bensard

April 11, 2018

Problem 1:

Consider the two slightly different set of data:

	$x = 0$	$x = 1$	Total
$y = 0$	2	2	4
$x = 1$	5	1	6
Total	7	3	10

Table 1: Data 1

	$x = 0$	$x = 1$	Total
$y = 0$	1	2	3
$x = 1$	5	2	7
Total	6	4	10

Table 2: Data 2

For the first set of data, we have $\widehat{OR_1} = \frac{2\cdot 1}{2\cdot 5} = \frac{1}{5}$ and $\widehat{RR_1} = \frac{2\cdot(2+1)}{2\cdot(2+5)} = \frac{3}{7}$.

For the second set of data, we have $\widehat{OR_2} = \frac{1\cdot 2}{2\cdot 5} = \frac{1}{5}$ and $\widehat{RR_2} = \frac{1\cdot(2+2)}{2\cdot(1+5)} = \frac{1}{3}$.

Note that $\widehat{OR_1} = \widehat{OR_2}$ but $\widehat{RR_1} \neq \widehat{RR_2}$. Therefore, from two different set of data giving similar Odds Ratio, we can get different Relative Risk. As a result, only provided the Odds Ratio, we cannot determine the Relative Risk.

Problem 2:

(a)

Let y be the learning impairment such that $y = \begin{cases} 0 \text{ if unimpaired} \\ 1 \text{ if impaired} \end{cases}$ and let x be the drinking during pregnancy such that $x = \begin{cases} 0 \text{ if no drinking} \\ 1 \text{ if drinking} \end{cases}$.

Thus, according to the data, we have $\widehat{OR}(y = 1|x = 0 \text{ or } x = 1) = \frac{172\cdot 24}{88\cdot 16} = \frac{129}{44} \simeq 2.932$.

(b)

The current data provides only one predictor to predict the learning impairment of children, therefore we can use Woolf's method in that case. Thus, we get the following 95% confidence interval for $\widehat{OR}$:

$$
\begin{aligned}
CI_{95\%} &= e^{ln(\widehat{OR})\pm z^*\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}} \\
&= e^{ln(2.932)\pm 1.96\sqrt{\frac{1}{24}+\frac{1}{88}+\frac{1}{16}+\frac{1}{172}}} \\
&= [1.481, 5.803]
\end{aligned}
$$

(c)

The previous model is certainly too simplistic. A lot of factors related to drinking might have an effect on the learning impairment of children and should be taken into consideration. Factors like being financially unstable, dealing with depression, being a single mom or any other negative factors that can get a pregnant mother to drink alcohol consistently might as well have a negative impact on the child after it is born, without taking the alcohol effect on the baby's metabolism into account. Therefore, we should not only look at the biological effect of alcohol consumption on the baby but look at all the co-factors that could possibly affect the child's learning abilities.

Problem 3:

(a)

```r
dat = read.csv('C:/Users/Louis/Documents/UPMC/M1/Spring 2018/MATH 535/Homework 4/hw4.csv',
               header=T, stringsAsFactors = FALSE)

#cleaning up data to make it usable
data = dat[,-1]
data[data$x3 =="yes",4]=1
data[data$x3 =="no",4]=0

maxit = 50; tolerr=1e-6; tolgrad=1e-9
beta_0 = c(0,0,0,0)

beta_MLE = newton(data, beta_0, maxit, tolerr, tolgrad)
```

```
##
## Iteration   Halving    log-likelihood    ||gradient||
##     1                     -138.6294        4.3e+02
##     1          0           -94.1385        8.3e+01
## ----------------------------------------------------------
##     2                      -94.1385        1.0e+02
##     2          0           -88.8735        2.0e+01
## ----------------------------------------------------------
## ...
## ...
## ...
##     4                      -88.3963        2.7e+00
##     4          0           -88.3899        3.4e-02
## ----------------------------------------------------------
##     5                      -88.3899        4.1e-02
##     5          0           -88.3899        7.9e-06
## ----------------------------------------------------------
```

```
##     6                       -88.3899       9.7e-06
##     6          0            -88.3899       5.1e-13
## ----------------------------------------------------------
## ----------------------------------------------------------
```

```r
print(beta_MLE)
```

```
##             [,1]
##     6.57336800
## x1 -0.66750844
## x2 -0.08735332
## x3 -0.28875755
```

Therefore, the estimate of the coefficients β of our generalized linear logistic model using Newton-Raphson algorithm is $\hat{\beta} = (6.5733, -0.6675, -0.0874, -0.2888)^T$.

(b)

```r
model = glm(y ~ x1+x2+x3, family = "binomial", data=data)
summary(model)
```

```
##
## Call:
## glm(formula = y ~ x1 + x2 + x3, family = "binomial", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0747  -0.6775  -0.2517   0.7180   2.7170
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.57337    3.01849   2.178   0.0294 *
## x1          -0.66751    0.09711  -6.873 6.27e-12 ***
## x2          -0.08735    0.96985  -0.090   0.9282
## x31         -0.28876    0.38474  -0.751   0.4529
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 272.12  on 199  degrees of freedom
## Residual deviance: 176.78  on 196  degrees of freedom
## AIC: 184.78
##
## Number of Fisher Scoring iterations: 5
```

The glm() function does produce similar coefficients as our algorithm.

(c)

Let $\Delta_1 = 5$ be the increase in x_1, and $\hat{\beta}$ be the estimate of the coefficients. Then, Odds Ratio resulting from this 5-unit increase will be:

$$
\begin{aligned}
\widehat{OR} &= e^{\Delta_1 \hat{\beta}_1} \\
&= e^{5 \cdot (-0.6675)} \\
&= 0.0355
\end{aligned}
$$