STAC51A2

Yulun Wu

2023-02-06

Q1

(a)

```
set.seed(1004912785)
n_q1 = 10
N_q1 = 100
z = qnorm(1-(0.05/2),0,1)
count=0 # counter to count #(intervals contain the true odd ratio)
# Generate random data
x_q1 = rmultinom(n_q1,N_q1,c(0.2,0.3,0.3,0.2))
x_q1
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]   19   17   17   12   24   14   15   18   23    27
## [2,]   34   34   29   38   32   24   28   30   20    22
## [3,]   22   25   31   33   24   38   29   31   41    32
## [4,]   25   24   23   17   20   24   28   21   16    19
```

```
odd_ratio_p = 0.2*0.2/(0.3*0.3) # true odd ratio
cat("True odd ratio is",odd_ratio_p)
```

```
## True odd ratio is 0.4444444
```

```
for (i in 1:n_q1) {
  x = as.table(matrix(x_q1[,i],nrow = 2,byrow = T))
  cat(i,"th table\n")
  print(x)
  odd_ratio_hat = x[1,1]*x[2,2]/(x[2,1]*x[1,2]) # estimated odd ratio
  cat(i,"th estimated odd ratio:",odd_ratio_hat,"\n")
  selog_ort = sqrt((1/x[1,1])+(1/x[1,2])+(1/x[2,1])+(1/x[2,2])) #SE(log(odd ratio))
  # Confidence interval
  CIl = exp(log(odd_ratio_hat)-z*selog_ort)
  CIu = exp(log(odd_ratio_hat)+z*selog_ort)
  CI=cbind(CIl,CIu)
  cat(i,"th 95% large sample CI:",CI,"\n")
  if ((CIl<=odd_ratio_p)&(CIu>=odd_ratio_p)) {
    count = count+1
  }
}
```

```
## 1 th table
##    A  B
## A 19 34
## B 22 25
## 1 th estimated odd ratio: 0.6350267
## 1 th 95% large sample CI: 0.2847246 1.416312
## 2 th table
##    A  B
## A 17 34
## B 25 24
## 2 th estimated odd ratio: 0.48
## 2 th 95% large sample CI: 0.2139849 1.076712
## 3 th table
##    A  B
## A 17 29
## B 31 23
## 3 th estimated odd ratio: 0.4349277
## 3 th 95% large sample CI: 0.1942886 0.973614
## 4 th table
##    A  B
## A 12 38
## B 33 17
## 4 th estimated odd ratio: 0.1626794
## 4 th 95% large sample CI: 0.06789389 0.3897935
## 5 th table
##    A  B
## A 24 32
## B 24 20
## 5 th estimated odd ratio: 0.625
## 5 th 95% large sample CI: 0.2822001 1.384213
## 6 th table
##    A  B
## A 14 24
## B 38 24
## 6 th estimated odd ratio: 0.3684211
## 6 th 95% large sample CI: 0.160004 0.848317
## 7 th table
##    A  B
## A 15 28
## B 29 28
## 7 th estimated odd ratio: 0.5172414
## 7 th 95% large sample CI: 0.2291304 1.167626
## 8 th table
##    A  B
## A 18 30
## B 31 21
## 8 th estimated odd ratio: 0.4064516
## 8 th 95% large sample CI: 0.1816877 0.9092685
## 9 th table
##    A  B
## A 23 20
## B 41 16
## 9 th estimated odd ratio: 0.4487805
## 9 th 95% large sample CI: 0.1952227 1.031663
## 10 th table
##    A  B
## A 27 22
## B 32 19
## 10 th estimated odd ratio: 0.7286932
## 10 th 95% large sample CI: 0.327604 1.62084
```

```
cat(count,"of intervals contain the true odd ratio",odd_ratio_p)
```

```
## 9 of intervals contain the true odd ratio 0.4444444
```

(b)

```
set.seed(1004912785)
n = 1000000
N = 100
z = qnorm(1-(0.05/2),0,1)
# Generate random data
x_q1 = rmultinom(n,N,c(0.2,0.3,0.3,0.2))
x_q1 = replace(x_q1,x_q1==0,0.5)
odd_ratio_p = 0.2*0.2/(0.3*0.3) # true odd ratio
odd_ratio_hat = x_q1[1,]*x_q1[4,]/(x_q1[2,]*x_q1[3,]) # estimated odd ratio
#SE(log(odd ratio))
selog_ort = sqrt((1/x_q1[1,])+(1/x_q1[2,])+(1/x_q1[3,])+(1/x_q1[4,]))
# Confidence interval
CIl = exp(log(odd_ratio_hat)-z*selog_ort)
CIu = exp(log(odd_ratio_hat)+z*selog_ort)
CI = cbind(CIl,CIu)
# Boolean to measure whether intervals contain the true odd ratio
includeBool = (CI[,1]<=odd_ratio_p)&(CI[,2]>=odd_ratio_p)
proportion = sum(includeBool*1)/n
cat(100*proportion,"% of intervals contain the true odd ratio",odd_ratio_p)
```

```
## 94.9487 % of intervals contain the true odd ratio 0.4444444
```

The proportion of intervals containing true odd ratio is around 95%, which is the level of the confidence intervals we are testing with. This means the coverage probability of large sample confidence interval of odd ratio is around 100*(1-α) if n and N is large.

(c)

```
set.seed(1004912785)
n = 1000000
N = 15
z = qnorm(1-(0.05/2),0,1)
# Generate random data
x_q1 = rmultinom(n,N,c(0.2,0.3,0.3,0.2))
x_q1 = replace(x_q1,x_q1==0,0.5)
odd_ratio_p = 0.2*0.2/(0.3*0.3) # true odd ratio
odd_ratio_hat = x_q1[1,]*x_q1[4,]/(x_q1[2,]*x_q1[3,]) # estimated odd ratio
#SE(log(odd ratio))
selog_ort = sqrt((1/x_q1[1,])+(1/x_q1[2,])+(1/x_q1[3,])+(1/x_q1[4,]))
# Confidence interval
CIl = exp(log(odd_ratio_hat)-z*selog_ort)
CIu = exp(log(odd_ratio_hat)+z*selog_ort)
CI = cbind(CIl,CIu)
# Boolean to measure whether intervals contain the true odd ratio
includeBool = (CI[,1]<=odd_ratio_p)&(CI[,2]>=odd_ratio_p)
```

```
proportion = sum(includeBool*1)/n
cat(100*proportion,"% of intervals contain the true odd ratio",odd_ratio_p)
```

```
## 97.159 % of intervals contain the true odd ratio 0.4444444
```

The proportion of intervals containing true odd ratio is above 95%, which is the level of the confidence intervals we are testing with. This means the coverage probability of large sample confidence interval of odd ratio is around or above 100*(1-α) if n is large. And the coverage probability tends to be higher when N is smaller.

Q3

##(a)

```
# Our data
data = matrix(c(21,2,15,3),2,2,byrow=T)
# p-value for one sided test, P(X>=n11)
pvalue = sum(dhyper(21:23,m=sum(data[1,]),n=sum(data[2,]),k=sum(data[,1])))
pvalue
```

```
## [1] 0.3808337
```

P-value is 0.3808337, so we fail to reject the null hypothesis that surgery and radiation therapy have no difference in controlling cancer at α=0.05.

##(b)

```
# density when X=21 (P(X=21))
density21 = dhyper(21,m=sum(data[1,]),n=sum(data[2,]),k=sum(data[,1]))
# All densities (P(X=i),i=0,...23)
densityall = dhyper(0:23,m=sum(data[1,]),n=sum(data[2,]),k=sum(data[,1]))
# Boolean value of whether sensity is nore than P(X=21) or not
boolcompare = densityall<=density21
# p-value for two sided test,sum(density no more like obs)
p_value = sum(boolcompare*densityall)
p_value
```

```
## [1] 0.6384258
```

P-value is 0.6384258, so we fail to reject the null hypothesis that surgery and radiation therapy have no difference in controlling cancer at α=0.05.

Q4

##(a) Equations are in writing file

```
data = matrix(c(154,180,104,132,126,131),3,2) # obs
n=sum(data) # n++
n_rowsum = as.matrix(rowSums(data)) # ni+
n_colsum = as.matrix(colSums(data)) # n+j
expected = n_rowsum%*%t(n_colsum)/n # expected
cat("Expected count is:\n")
```

```
## Expected count is:
```

```
expected
```

```
##           [,1]     [,2]
## [1,] 151.4728 134.5272
## [2,] 162.0653 143.9347
## [3,] 124.4619 110.5381
```

```
# Pearson X^2 test
Xsq = sum(((data-expected)^2)/expected) # test statistics for Pearson X^2
cat("Test statistics for Pearson X^2 is",Xsq,"\n")
```

```
## Test statistics for Pearson X^2 is 11.46082
```

```
pchisq(Xsq,df=2,lower.tail = F) # p-value for X^2
```

```
## [1] 0.003245752
```

```
# G^2 likelihood ratio test of independence
Gsq = sum(2*(data*log(data/expected))) # test statistics for likelihood ratio test of independence G^2
cat("Test statistics for likelihood ratio test of independence G^2 is",Gsq,"\n")
```

```
## Test statistics for likelihood ratio test of independence G^2 is 11.4775
```

```
pchisq(Gsq,df=2,lower.tail = F) # p-value for G^2
```

```
## [1] 0.003218785
```

Both p-value from Pearson X^2 test (0.003245752) and p-value from likelihood ratio test of independence G^2 (0.003218785) are less than 0.05, so we reject the null hypothesis that whether to support offshore drilling is independent from whether the voters are graduated from College at α=0.05.

##(b)

```
p1 = data[1,1]/n_colsum[1]
p2 = data[1,2]/n_colsum[2]
p=(data[1,1]+data[1,2])/n
test_stat = (p1-p2)/sqrt(p*(1-p)*(sum(1/n_colsum)))
cat("Test statistics for two-sample test of proportions is",test_stat,"\n")
```

```
## Test statistics for two-sample test of proportions is 0.3701737
```

```
p_value = 2*pnorm(test_stat,0,1,lower.tail = F)
cat("p-value for two-sample test of proportions is",p_value,"\n")
```

```
## p-value for two-sample test of proportions is 0.7112531
```

Since p-value=0.7112531>0.05, we fail to reject the null hypothesis that the proportion of college graduates supporting of offshore drilling equals to the proportion of non-college graduates supporting offshore drilling at α=0.05.

##(c) According to (a), we reject the null hypothesis that odd ratio $\theta = 1$. According to (b), we fail to reject the null hypothesis that $\pi_1 = \pi_2$, and $\pi_1 = \pi_2$ implies $\theta = 1$. This two conclusion don't agree each other. This disagreement is possible because p-value in (a) are computed based on all 6 cells and test statistics in (b) only computed from cell [1,1] and cell [1,2], so it is possible that only these twoo cells are independent.

##(d)

```
prow = n_rowsum/n # pi+
pcol = n_colsum/n # p+j
pmatrix = (1-prow)%*%t((1-pcol)) # 3 by 2 matrix of (1-pi+)(1-p+j)
s_residual = (data-expected)/sqrt(expected*pmatrix) # standardized residual
cat("Absolute value of standardized residual:\n")
```

```
## Absolute value of standardized residual:
```

```
abs(s_residual) # absolute value of standardized residual
```

```
##           [,1]      [,2]
## [1,] 0.3701737 0.3701737
## [2,] 2.5879807 2.5879807
## [3,] 3.1608057 3.1608057
```

Absolute value of standardized residual of cell [2,1],[2,2],[3,1],[3,2] are exceed 2, so H_0 fit badly on these cells. and they show strong evidence of association. Cell [1,1] and cell [1,2] have absolute standardized residual less than 2, so standardized residual of these two cells don't show strong evidence of association.

Q5

(a)

```
ECC = array(c(11,10,25,27,16,22,4,10,14,7,5,12,2,1,14,16,6,0,11,12,1,0,10,10,1,1,4,8,4,6,2,1),dim=c(2,2
ECC.margin = margin.table(ECC,c(1,2)) # Get marginal X,Y table
cat("Marginal X,Y table:\n")
```

```
## Marginal X,Y table:
```

```
ECC.margin
```

```
##          Response
## Treatment Success Failure
##   Drug         55      75
##   Control      47      96
```

```
odd_ratio_m = ECC.margin[1,1]*ECC.margin[2,2]/(ECC.margin[1,2]*ECC.margin[2,1]) # Calculate sample odd
cat("Sample odd ratio for marginal table is",odd_ratio_m,"\n")
```

```
## Sample odd ratio for marginal table is 1.497872
```

The marginal odd ratio for the marginal table for the Treatment and Response is 1.497872, this indicates that Drug group has larger probability of success than Control group.

(b)

```
odd_ratio_p = ECC[1,1,]*ECC[2,2,]/(ECC[1,2,]*ECC[2,1,]) # Sample odd ratio in partial table
cat("Sample odd ratio in partial table:\n")
```

```
## Sample odd ratio in partial table:
```

```
odd_ratio_p
```

```
##         1         2         3         4         5         6         7         8
## 1.1880000 1.8181818 4.8000000 2.2857143       Inf       Inf 2.0000000 0.3333333
```

Test the independence of the Treatment and Response using the marginal table of Treatment and Response ignore Center is not a good idea because the association between Treatment and Response in each Center can be different, and by using the marginal table, we are not make use of the information provided by Center. Also, there is chance that we observe Simpson's Paradox which means we get different result from partial table and marginal table, for example the sample partial odd ratio we get in the 8th partial table is 0.3333333 which indicates that in 8th Center, Drug group has less probability of success than Control group, this is contradict to the result we get from the marginal table.

(c)

```
expected11k = c()
varn11k = c()
nk=c()
for (i in 1:8) {
  nk = cbind(nk,sum(ECC[,,i]))
  n_rowsumk= rowSums(ECC[,,i])
  n_colsumk = colSums(ECC[,,i])
  expected11k = cbind(expected11k,n_rowsumk[1]%*%n_colsumk[1]/nk[i]) # expected value of cell 11i
  varn11k =  cbind(varn11k,prod(n_rowsumk,n_colsumk)/((nk[i]^2)*(nk[i]-1))) # variance of cell 11i
}
expected11k # expected value of cell (Drug, Success) for all Centers
```

```
##          [,1]     [,2] [,3]     [,4]     [,5]      [,6]      [,7]     [,8]
## [1,] 10.35616 14.61538 10.5 1.454545 3.517241 0.5238095 0.7142857 4.615385
```

```
varn11k # variance of cell (Drug, Success) for all Centers
```

```
##          [,1]     [,2]     [,3]      [,4]     [,5]      [,6]      [,7]
## [1,] 3.790955 2.468964 2.412162 0.7024793 1.195516 0.2494331 0.4238619
##           [,8]
## [1,] 0.6213018
```

```
CMH = sum(ECC[1,1,]-expected11k)/sqrt(sum(varn11k)) # CMH test statistic
cat("CMH test statistic:\n")
```

```
## CMH test statistic:
```

```
CMH
```

```
## [1] 2.52668
```

```
p_value = 2*pnorm(CMH,0,1,lower.tail = F) # p-value of CMH
cat("CMH value:\n")
```

```
## CMH value:
```

```
p_value
```

```
## [1] 0.01151463
```

$$
\begin{aligned}
CMH &= \frac{\sum_{i=1}^{k}(n_{11i} - \mathbb{E}[n_{11i}])}{\sqrt{\sum_{i=1}^{k} Var[n_{11i}]}} \\
&= \frac{(11-10.35616)+(16-14.61538)+(14-10.5)+(2-1.454545)}{\sqrt{3.790955+2.468964+2.412162+0.7024793+1.195516+0.2494331+0.4238619+0.6213018}} \\
&+ \frac{(6-3.517241)+(1-0.5238095)+(1-0.7142857)+(4-4.615385)}{\sqrt{3.790955+2.468964+2.412162+0.7024793+1.195516+0.2494331+0.4238619+0.6213018}}
\end{aligned} \tag{1}
$$

P-value of CMH test is 0.01151463<0.05, so we reject the H_0 that Treatment and Response are conditionally independent given Center at $\alpha = 0.05$.

(d)

```
nk
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
## [1,]   73   52   38   33   29   21   14   13
```

```
# Mantel-Haenszel's estimate of common odds ratio between Treatment and Response
odd_ratio_MH = sum(ECC[1,1,]*ECC[2,2,]/nk)/sum(ECC[2,1,]*ECC[1,2,]/nk)
cat("Mantel-Haenszel's estimate of common odds ratio between Treatment and Response is",odd_ratio_MH,"\
```

```
## Mantel-Haenszel's estimate of common odds ratio between Treatment and Response is 2.134549
```

$$\begin{aligned}\widehat{\theta}_{MH} &= \frac{\sum_{i=1}^{k} \frac{n_{11i}*n_{22i}}{T_i}}{\sum_{i=1}^{k} \frac{n_{12i}*n_{21i}}{T_i}} \\ &= \frac{\frac{11*27}{73}+\frac{6*10}{52}+\frac{14*12}{38}+\frac{2*16}{33}+\frac{6*12}{29}+\frac{1*10}{21}+\frac{1*8}{14}+\frac{4*1}{13}}{\frac{25*10}{73}+\frac{4*22}{52}+\frac{5*7}{38}+\frac{14*1}{33}+\frac{11*0}{29}+\frac{10*0}{21}+\frac{4*1}{14}+\frac{2*6}{13}}\end{aligned} \tag{2}$$

$\widehat{\theta}_{MH} = 2.134549$,

(e)

```
mantelhaen.test(ECC,correct = F)
```

```
##
##  Mantel-Haenszel chi-squared test without continuity correction
##
## data:  ECC
## Mantel-Haenszel X-squared = 6.3841, df = 1, p-value = 0.01151
## alternative hypothesis: true common odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.177590 3.869174
## sample estimates:
## common odds ratio
##          2.134549
```

This is the same as the result I computed manually.

Q6

(a)

```
mydata = data.frame(drinks = c(0,0.5,1.5,4,7),
absent = c(17066, 14464, 788, 126, 37),
present = c(48, 38, 5, 1, 1) )
mydata$total = with(mydata, absent + present)
mydata$proportion = with(mydata, present/total)
data.linear = glm(cbind(present,absent)~drinks,family=binomial(link="identity"),data=mydata)
summary(data.linear)
```

```
##
## Call:
## glm(formula = cbind(present, absent) ~ drinks, family = binomial(link = "identity"),
##     data = mydata)
```

```
##
## Deviance Residuals:
##       1        2        3        4        5
##  0.6564  -1.0492   0.8631   0.1302   0.8282
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.0025476  0.0003523   7.232 4.77e-13 ***
## drinks      0.0010872  0.0008324   1.306    0.192
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6.2020  on 4  degrees of freedom
## Residual deviance: 2.9795  on 3  degrees of freedom
## AIC: 25.606
##
## Number of Fisher Scoring iterations: 10
```

(b)

```
pi0 = 0.0025476+0.0010872*0 # pi(0)
pi0
```

```
## [1] 0.0025476
```

```
pi7 = 0.0025476+0.0010872*7 # pi(7)
pi7
```

```
## [1] 0.010158
```

```
relative_risk = pi0/pi7 # the relative risk comparing the two levels
cat("The relative risk comparing the two levels is",relative_risk,"\n")
```

```
## The relative risk comparing the two levels is 0.2507974
```

$$\pi(x) = 0.0025476 + 0.0010872x$$

. Interpretation: If mother's alcohol consumption is 0, then the chance that a baby has sex organ malformation is 0.0025476. If the alcohol consumption is increased by 1, then the chance that a baby has sex organ malformation will be increased by 0.0010872. The probabilities of malformation for the lowest and highest alcohol levels: $\pi(0) = 0.0025476$ and $\pi(7) = 0.010158$. The relative risk comparing the two levels in part (iii) is 0.2507974. The probability of baby has sex organ malformation case for mother's alcohol consumption is 0 is 0.2507974 times that for mother's alcohol consumption is 7.

(c)

```
CIl = 0.0010872-qnorm(1-0.1/2)*0.0008324 # lower CI
CIu = 0.0010872+qnorm(1-0.1/2)*0.0008324 # upper CI
CI = cbind(CIl,CIu) # Wald confidence interval
cat("Wald confidence interval:\n")
```

```
## Wald confidence interval:
```

```
CI
```

```
##                  CIl         CIu
## [1,] -0.0002819762 0.002456376
```

(d)

```
data.logit = glm(cbind(present,absent)~drinks,family=binomial(link="logit"),data=mydata)
summary(data.logit)
```

```
## 
## Call:
## glm(formula = cbind(present, absent) ~ drinks, family = binomial(link = "logit"), 
##     data = mydata)
## 
## Deviance Residuals: 
##       1        2        3        4        5  
##  0.5921  -0.8801   0.8865  -0.1449   0.1291  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  -5.9605     0.1154 -51.637   <2e-16 ***
## drinks        0.3166     0.1254   2.523   0.0116 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 6.2020  on 4  degrees of freedom
## Residual deviance: 1.9487  on 3  degrees of freedom
## AIC: 24.576
## 
## Number of Fisher Scoring iterations: 4
```

```
pi0 = exp(-5.9605+0.3166*0)/(1+exp(-5.9605+0.3166*0)) # pi(0)
pi0
```

```
## [1] 0.00257199
```

```
pi7 = exp(-5.9605+0.3166*7)/(1+exp(-5.9605+0.3166*7)) # pi(7)
pi7
```

```
## [1] 0.02310568
```

```
relative_risk = pi0/pi7 # the relative risk comparing the two levels
cat("The relative risk comparing the two levels is",relative_risk,"\n")
```

```
## The relative risk comparing the two levels is 0.1113142
```

```
odd_ratio = pi0*(1-pi7)/(pi7*(1-pi0)) # odd ratio of level 0 and 7
odd_ratio
```

```
## [1] 0.1090226
```

```
cat("The odd ratio of level 0 and 7 is",odd_ratio,"\n")
```

```
## The odd ratio of level 0 and 7 is 0.1090226
```

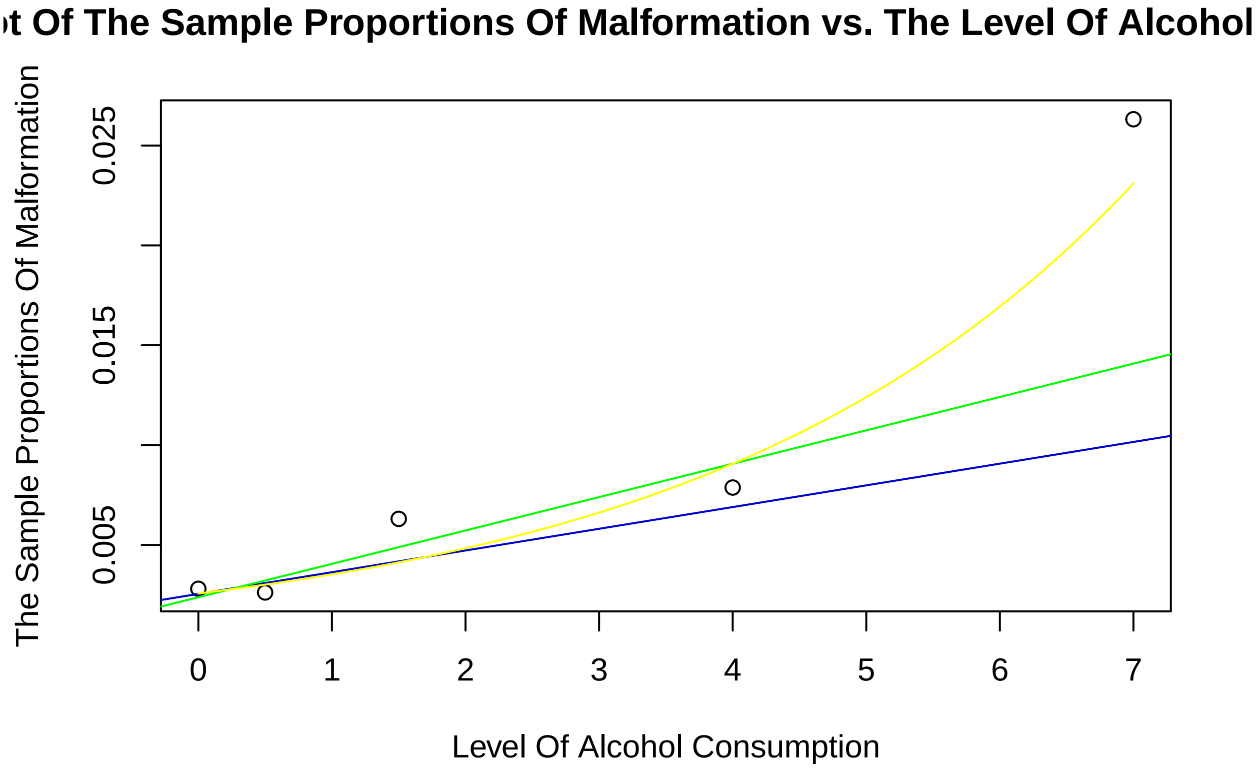
$$\widehat{\pi}(x) = \frac{e^{-5.9605+0.3166x}}{1+e^{-5.9605+0.3166x}}$$

. Interpretation: When the mother's alcohol consumption is 0, the probability that the baby has sex organ malformation is $\frac{e^{-5.9605}}{1+e^{-5.9605}}$. $log(odds) = -5.9605 + 0.3166x$, the odds increases by a factor of $e^{0.3166}$ whenever the mother's alcohol consumption increases by 1. The probabilities of malformation for the lowest and highest alcohol levels: $\pi(0) = 0.00257199$ and $\pi(7) = 0.02310568$. The relative risk comparing the two levels in part (iii) is 0.1113142. The odds ratio of malformations for alcohol levels 7 vs. 0 is 0.1090226.

(e)

```
# Make grouped data ungrouped
data.UG = mydata[rep(1:nrow(mydata),mydata$present),cbind("present","drinks")]
data.UG$present = 1
data.UGtemp = mydata[rep(1:nrow(mydata),mydata$absent),cbind("absent","drinks")]
data.UGtemp$absent=0
colnames(data.UGtemp)[1]="present"
data.UG = rbind(data.UG,data.UGtemp)

plot(mydata$drinks, mydata$present/rowSums(cbind(mydata$present,mydata$absent)), main="Scatterplot Of Th
   xlab="Level Of Alcohol Consumption", ylab="The Sample Proportions Of Malformation")
abline(data.linear,col="blue3") # the fitted line by the ML method from part (a)
abline(lm(present~drinks,data = data.UG),col="green")
curve(exp(-5.9605+0.3166*x)/(1+exp(-5.9605+0.3166*x)), col="yellow",add=T)
```



The slope differ so much for the two straight lines is Least Square method assume constant variability in all points, which is not true in our case.