# HW3_yc4384_Cynthia

Yangyang Chen

2024-02-26

## Problem 1

**(a) Fit a prospective model to the data to study the relation between alcohol consumption, age, and disease (model age as a continuous variable taking values 25, 35, 45, 55, 65, and 75). Interpret the result.**

Using logistics regression model to fit data from prospective study:

$$log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 X_{alc} + \beta_2 X_{age}$$

```
# load data
age = seq(from = 25, to = 75, by = 10) |>
  rep(2)
case = c(1, 4, 25, 42, 19, 5, 0, 5, 21, 34, 36, 8)
control = c(9, 26, 29, 27, 18, 0, 106, 164, 138, 139, 88, 31)
alc = c(rep(1,6), rep(0, 6))
resp = cbind(case, control)

# Model fitting using logit link
glm_logit=glm(resp ~ alc + age, family=binomial(link='logit'))
summary(glm_logit)
```

```
##
## Call:
## glm(formula = resp ~ alc + age, family = binomial(link = "logit"))
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.023449   0.418224 -12.011   <2e-16 ***
## alc          1.780000   0.187086   9.514   <2e-16 ***
## age          0.061579   0.007291   8.446   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 211.608  on 11  degrees of freedom
## Residual deviance:  31.932  on  9  degrees of freedom
## AIC: 78.259
##
## Number of Fisher Scoring iterations: 4
```

Hence, the logistics regression model is:

$$log(\frac{\pi}{1-\pi}) = -5.02 + 1.78X_{alc} + 0.06 * X_{age}$$

Interpretation:

- The model suggests a significant relationship between esophageal cancer, daily alcohol consumption adjusted and age.
- $\beta_1$: the log odds ratio of having the disease among heavy drinkers is 1.78 times the odds odds ratio of non-heavy drinkers, keeping age fixed.
- $exp(\beta_1)$: odds ratio for the association between disease and alcohol consumption, holding age constant.
- $\beta_2$: the log odds ratio of having the disease will increase by 0.06 for every unit increment in age, keeping alcohol consumption fixed.
- $exp(\beta_2)$: the odds ratio for the association between disease and age, holding alcohol consumption constant.
- This model appears to fit the data well, as indicated by the significant coefficients and the reduction in deviance from the null model to the fitted model.

(b)

```
age = c(1:6) |>
  factor()
ind = dummy.code(age)
grp1 = rep(ind[,1],2)
grp2 = rep(ind[,2],2)
grp3 = rep(ind[,3],2)
grp4 = rep(ind[,4],2)
grp5 = rep(ind[,5],2)
grp6 = rep(ind[,6],2)

M_0 = glm(resp ~ grp1 + grp2 + grp3 + grp4 + grp5 + grp6, family = binomial(link = 'logit'))
summary(M_0)
```

```
##
## Call:
## glm(formula = resp ~ grp1 + grp2 + grp3 + grp4 + grp5 + grp6,
##     family = binomial(link = "logit"))
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.86904    0.33043  -2.630 0.008537 **
## grp1        -3.87589    1.05728  -3.666 0.000246 ***
## grp2        -2.18076    0.47493  -4.592 4.39e-06 ***
## grp3        -0.42031    0.37001  -1.136 0.255977
## grp4         0.08778    0.35828   0.245 0.806445
## grp5         0.21293    0.36986   0.576 0.564812
## grp6              NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 211.608  on 11  degrees of freedom
```

```
## Residual deviance:  90.563  on  6  degrees of freedom
## AIC: 142.89
##
## Number of Fisher Scoring iterations: 6
```

```
dev_m0 = residuals(M_0, type = "deviance")^2 |> sum()

M_1 = glm(resp ~ alc + grp1 + grp2 + grp3 + grp4 + grp5 + grp6, family = binomial(link = 'logit'))
summary(M_1)
```

```
##
## Call:
## glm(formula = resp ~ alc + grp1 + grp2 + grp3 + grp4 + grp5 +
##     grp6, family = binomial(link = "logit"))
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.092158   0.344216  -3.173 0.001509 **
## alc          1.669890   0.189602   8.807  < 2e-16 ***
## grp1        -3.962190   1.065035  -3.720 0.000199 ***
## grp2        -2.419896   0.491328  -4.925 8.43e-07 ***
## grp3        -0.763428   0.389837  -1.958 0.050192 .
## grp4        -0.248700   0.376735  -0.660 0.509161
## grp5         0.004692   0.387043   0.012 0.990328
## grp6               NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 211.608  on 11  degrees of freedom
## Residual deviance:  11.041  on  5  degrees of freedom
## AIC: 65.369
##
## Number of Fisher Scoring iterations: 5
```

```
dev_m2 = residuals(M_1, type = "deviance")^2 |> sum()
```

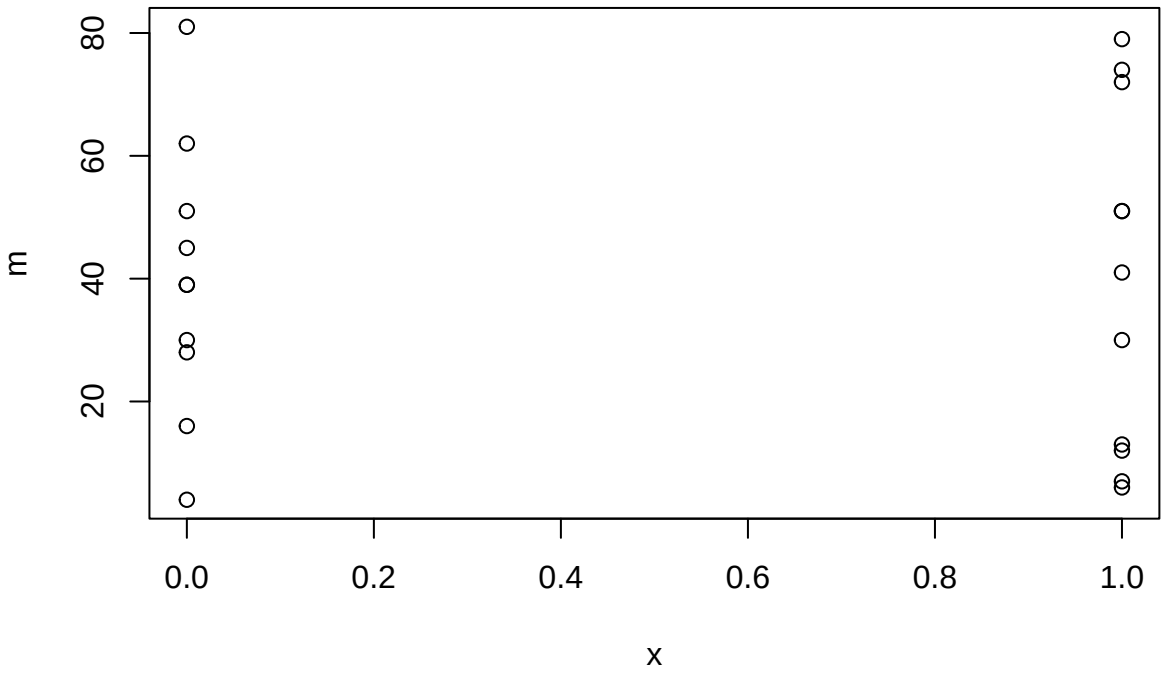
## Problem 2

**(a) Fit a logistic regression model to study the relation between germination rates and different types of seed and root extract. Interpret the result.**

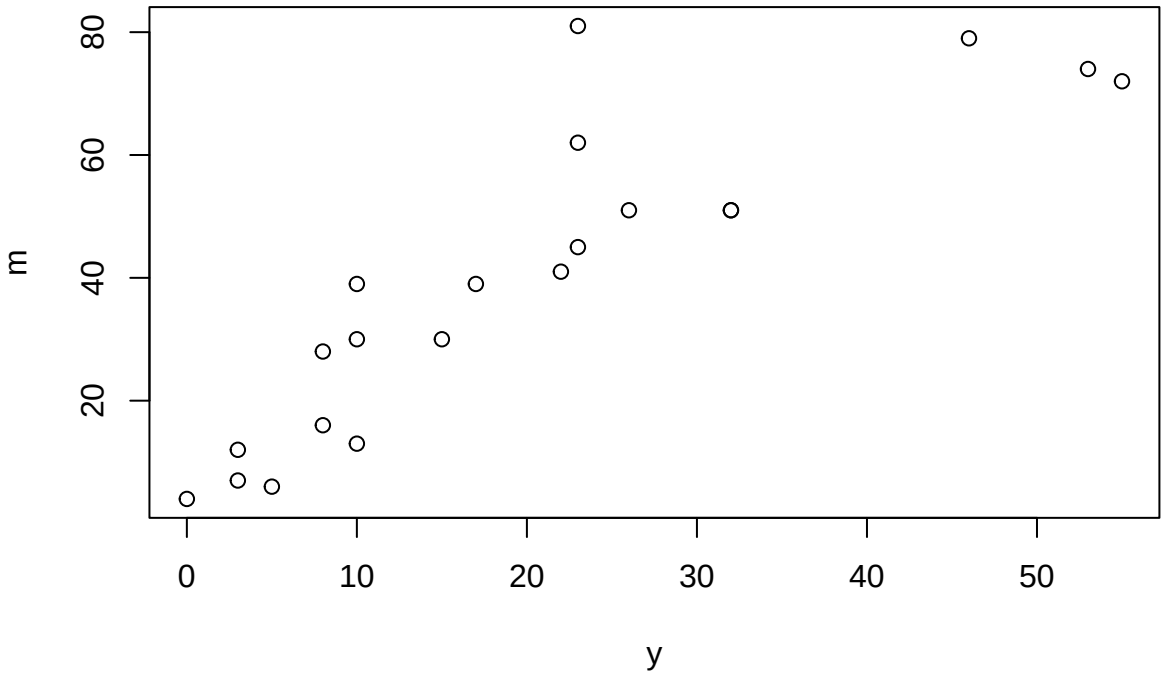
Let $Y_i$ denote the number of seeds germinates among $m_i$ seeds with the $ith$ covariate pattern. The logistic regression model:

$$log(\frac{\pi}{1-\pi}) = \beta_1 + \beta_2 x_i;$$
$$Y_i \sim Bin(n_i, \pi_i),\ i = 1, ..., m,$$

```
# input data
x=c(rep(0,10),rep(1,11))
y=c(10,23,23,26,17,8,10,8,23,0,5,53,55,32,46,10,3,22,15,32,3) # survive=1
m=c(39,62,81,51,39,16,30,28,45,4,6,74,72,51,79,13,12,41,30,51,7)
data=data.frame(x,y,m)
plot(x,m)
```




```
plot(y,m)
```




```
summary(m-y) ## m >= y
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    9.00   19.00   19.38   22.00   58.00
```

```
# fit binomial (logistic) without dispersion
none.disp=glm(cbind(y,m-y)~x, family=binomial(link='logit'))
summary(none.disp)
```

```
##
## Call:
## glm(formula = cbind(y, m - y) ~ x, family = binomial(link = "logit"))
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.5122     0.1039  -4.927 8.34e-07 ***
## x             1.0574     0.1438   7.353 1.93e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 98.719  on 20  degrees of freedom
## Residual deviance: 42.751  on 19  degrees of freedom
## AIC: 123.35
##
## Number of Fisher Scoring iterations: 3
```

```
G.stat=sum(residuals(none.disp,type='pearson')^2) # pearson chisq
G.stat
```

```
## [1] 41.22615
```

The binomial logistic without dispersion was fitted with the following results:

$$b_1 = -0.5122,\ standard\ error(b_1) = 0.1039;$$
$$b_2 = 1.0574,\ standard\ error(b_2) = 0.1438;$$
$$Pearson - \chi^2 statistic: X^2 = \sum X_i^2 = 41.226\ and\ Deviance\ D = \sum d_i^2 = 42.751.$$

```
# goodness of fit
pval=1-pchisq(none.disp$deviance,21-3)
pval # bad fit, reject the fitting
```

```
## [1] 0.0008677067
```

Comparing $X^2$ and $D$ with $\chi^2(19)$, we concluded that the model appears to fit bad.

**(b) Is there over dispersion? If so, what is the estimate of dispersion parameter? Update your model and reinterpret the result.**

Estimating the dispersion parameter by following two methods:

First,

$$\hat{\phi} = G_0/(n-p),$$

where

$$G_0 = \sum_{i=1}^{n} \frac{(y_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1-\hat{\pi}_i)\phi} \sim \chi^2(n-p)$$

is the generalized Pearson $\chi^2$ from the original model fitting without over-dispersion.

Second,

$$\hat{\phi} = \frac{D_0}{n-p}$$

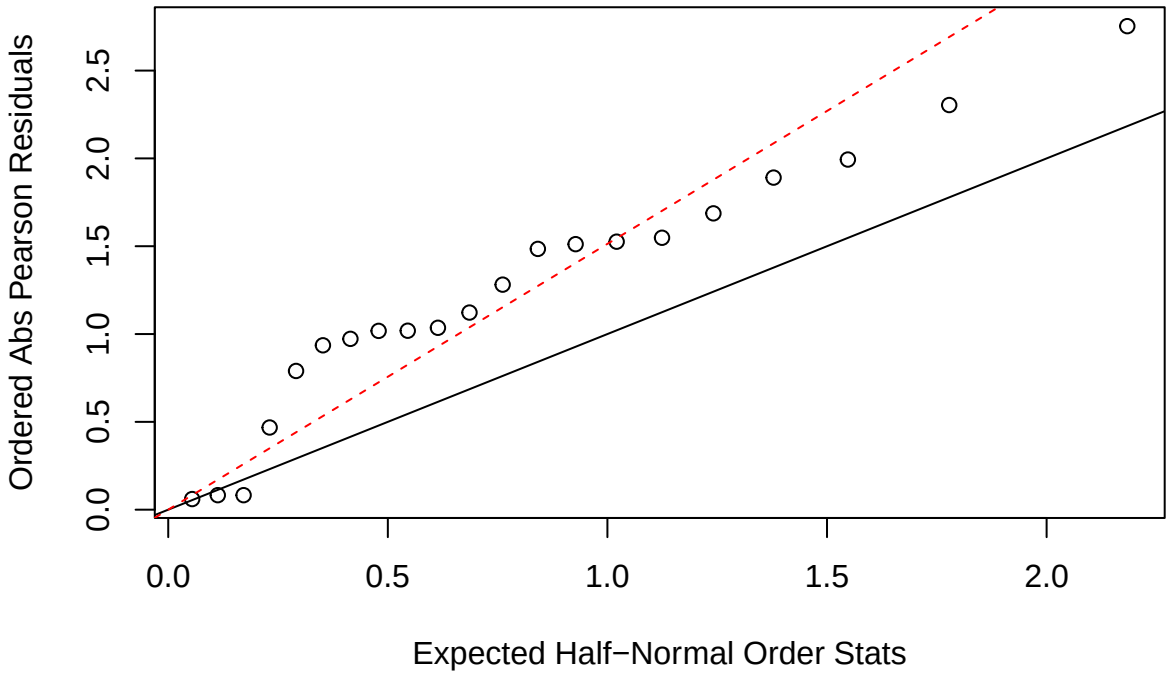
```r
# calc dispersion para in 2 methods
# the first method
phi=G.stat/(21-3)
phi
```

```
## [1] 2.290341
```

```r
# the second method
tilde.phi=none.disp$deviance/none.disp$df.residual
tilde.phi # similar to the one estimated from pearson chisq
```

```
## [1] 2.250045
```

```r
# test over-dispersion (half normal plot)
res=residuals(none.disp,type='pearson')
plot(qnorm((21+1:21+0.5)/(2*21+1.125)),sort(abs(res)),xlab='Expected Half-Normal Order Stats',ylab='Ord
abline(a=0,b=1)
abline(a=0,b=sqrt(phi),lty=2, col = 'red')
```




Therefore, there exists over-dispersion in our model and the estimate of dispersion parameter: $\hat{\phi} = 2.1697$.

Next, we updated regression model.

```r
# fit model with constant over-dispersion
summary(none.disp,dispersion=phi)
```

```
##
## Call:
## glm(formula = cbind(y, m - y) ~ x, family = binomial(link = "logit"))
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.5122     0.1573  -3.256  0.00113 **
## x             1.0574     0.2176   4.859 1.18e-06 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 2.290341)
##
##     Null deviance: 98.719  on 20  degrees of freedom
## Residual deviance: 42.751  on 19  degrees of freedom
## AIC: 123.35
##
## Number of Fisher Scoring iterations: 3
```

The binomial logistic with dispersion was fitted with the following results:

$$b_1 = -0.5122,\ standard\ error(b_1) = 0.1531;$$
$$b_2 = 1.0574,\ standard\ error(b_2) = 0.2118;$$
$$Pearson - \chi^2 statistic : X^2 = \sum X_i^2 = 41.226\ and\ Deviance\ D = \sum d_i^2 = 42.751.$$

**(c) What is a plausible cause of the over dispersion?**