hw3 Logistic regression case study

Cary Ni

2023-02-24

Problem 1

```
# input observations
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)
age = rep(c(25, 35, 45, 55, 65, 75), 2)
exposure = c(rep(1, 6), rep(0, 6))
pros_model = glm(cbind(case, control)~exposure + age, family = binomial(link = "logit"))
summary(pros_model)
```

```
##
## Call:
## glm(formula = cbind(case, control) ~ exposure + age, family = binomial(link = "logit"))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.59974  -1.72957   0.06822   1.19015   1.50808
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.023449   0.418224 -12.011   <2e-16 ***
## exposure     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
```

(a) Interpretation: The group with exposure (daily alcohol consumption over 80g) has 1.2057479 times odds of developing esophageal cancer compared to the group without exposure (daily alcohol consumption below 80g) holding age group fixed. Every one year increase in age will lead to 1.0635156 times the odds of developing esophageal cancer holding exposure presence fixed.

```
# turn the age into dummy variables
dummy_age = dummy_cols(age)[-1]
colnames(dummy_age) = LETTERS[1:6]
# fit the model with only age
model_1 = glm(cbind(case, control)~data.matrix(dummy_age), family = binomial(link = "logit"))
summary(model_1)
```

```
##
## Call:
## glm(formula = cbind(case, control) ~ data.matrix(dummy_age),
##     family = binomial(link = "logit"))
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -3.477  -1.299   0.368   2.481   5.028
##
## Coefficients: (1 not defined because of singularities)
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -0.86904    0.33043  -2.630 0.008537 **
## data.matrix(dummy_age)A -3.87589    1.05728  -3.666 0.000246 ***
## data.matrix(dummy_age)B -2.18076    0.47493  -4.592 4.39e-06 ***
## data.matrix(dummy_age)C -0.42031    0.37001  -1.136 0.255977
## data.matrix(dummy_age)D  0.08778    0.35828   0.245 0.806445
## data.matrix(dummy_age)E  0.21293    0.36986   0.576 0.564812
## data.matrix(dummy_age)F       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
```

```
# fit the model with exposure added
model_2 = glm(cbind(case, control)~data.matrix(dummy_age) + exposure, family = binomial(link = "logit")
summary(model_2)
```

```
##
## Call:
## glm(formula = cbind(case, control) ~ data.matrix(dummy_age) +
##     exposure, family = binomial(link = "logit"))
##
## Deviance Residuals:
##        1         2         3         4         5         6         7         8
##  0.96641  -0.05538   0.13652   0.45905  -1.59342   2.11053  -1.16129   0.04747
##        9        10        11        12
## -0.11628  -0.35391   0.96513  -0.67850
##
## Coefficients: (1 not defined because of singularities)
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -1.092158   0.344216  -3.173 0.001509 **
## data.matrix(dummy_age)A -3.962190   1.065035  -3.720 0.000199 ***
## data.matrix(dummy_age)B -2.419896   0.491328  -4.925 8.43e-07 ***
## data.matrix(dummy_age)C -0.763428   0.389837  -1.958 0.050192 .
## data.matrix(dummy_age)D -0.248700   0.376735  -0.660 0.509161
## data.matrix(dummy_age)E  0.004692   0.387043   0.012 0.990328
## data.matrix(dummy_age)F        NA         NA      NA       NA
## exposure                 1.669890   0.189602   8.807  < 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:  11.041  on  5  degrees of freedom
## AIC: 65.369
##
## Number of Fisher Scoring iterations: 5
```

```
sum(residuals(model_2, type = "pearson")^2)
```

```
## [1] 9.319711
```

```
anova(model_1, model_2)
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(case, control) ~ data.matrix(dummy_age)
## Model 2: cbind(case, control) ~ data.matrix(dummy_age) + exposure
##   Resid. Df Resid. Dev Df Deviance
## 1         6     90.563
## 2         5     11.041  1   79.522
```

```
# get the p value
1-pchisq(79.522, 1)
```

```
## [1] 0
```

(b) It can be seen that the smaller model use age group as the only predictor which is nested by the larger model which use both age group and exposure status as covariates. The nested models are compared based on deviance, which follows chi-squared distribution with df = 1 and statistic = 79.522 with p value close to 0 (< 0.000001). Therefore, the model coefficient β for exposure status is not 0 while the null hypothesis is rejected, which provides the evidence for the association between alchol consumption and esophageal cancer.

Problem 2

```r
# number of seed that germinates
y = c(10, 23, 23, 26, 17, 5, 53, 55, 32, 46, 10, 8, 10, 8, 23, 0, 3, 22, 15, 32, 3)
# total number of seed
m = c(39, 62, 81, 51, 39, 6, 74, 72, 51, 79, 13, 16, 30, 28, 45, 4, 12, 41, 30, 51, 7)
# set 0.a75 as 1
seed = c(rep(1, 11), rep(0, 10))
# set bean as 1
root = c(rep(1, 5), rep(0, 6), rep(1, 5), rep(0, 5))

# fit model without dispersion
model_3 = glm(cbind(y, m-y)~seed + root, family = binomial(link = "logit"))
summary(model_3)
```

```
##
## Call:
## glm(formula = cbind(y, m - y) ~ seed + root, family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3919  -0.9949  -0.3744   0.9831   2.4766
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.3643     0.1428   2.550   0.0108 *
## seed          0.2705     0.1547   1.748   0.0804 .
## root         -1.0647     0.1442  -7.383 1.55e-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: 39.686  on 18  degrees of freedom
## AIC: 122.28
##
## Number of Fisher Scoring iterations: 4
```

```r
# check model fit with chisq
1-pchisq(39.686, 18)
```

```
## [1] 0.00230269
```

```r
# get person statistic G
sum(residuals(model_3, type = "pearson")^2)
```

```
## [1] 38.31062
```

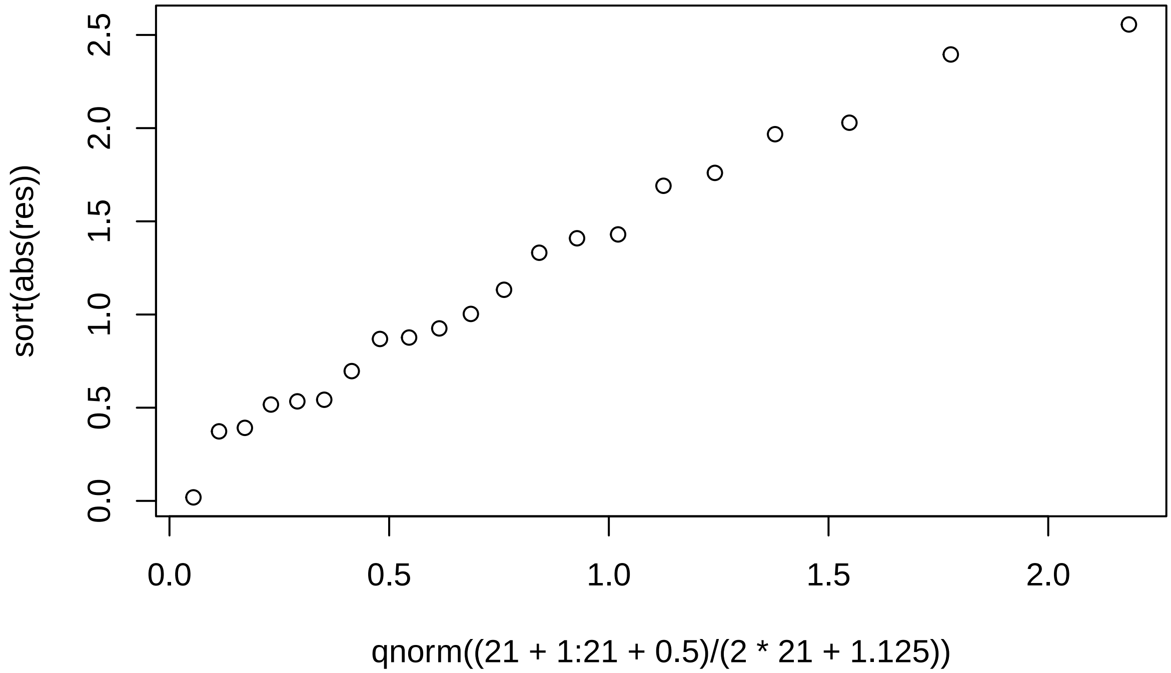
```r
# get dispersion factor from pearson G
sum(residuals(model_3, type = "pearson")^2)/18
```

```
## [1] 2.128368
```

```r
# get dispersion factor from residual deviance
39.686/18
```

```
## [1] 2.204778
```

```r
# half normal plot to check over-dispersion
res = residuals(model_3, type = "pearson")
plot(qnorm((21 + 1:21 + 0.5)/(2*21 + 1.125)), sort(abs(res)))
```



```r
# fit model with constant dispersion factor
summary(model_3, dispersion = 2.128)
```

```
##
## Call:
## glm(formula = cbind(y, m - y) ~ seed + root, family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3919  -0.9949  -0.3744   0.9831   2.4766
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.3643     0.2083   1.748   0.0804 .
## seed           0.2705     0.2257   1.198   0.2308
## root         -1.0647     0.2104  -5.061 4.17e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 2.128)
##
##     Null deviance: 98.719  on 20  degrees of freedom
## Residual deviance: 39.686  on 18  degrees of freedom
## AIC: 122.28
##
## Number of Fisher Scoring iterations: 4
```

```
# check model fit with chisq
1-pchisq(39.686/2.128, 18)
```

```
## [1] 0.4137027
```

(a) The odds of germination for O.aegyptiaca 73 seed in cucumber extract is 1.4333294. The odds of germination for O.aegyptiaca 75 seed is 1.3099645 times the odds for O.aegyptiaca 75 seed holding root extract fixed. The odds of germination in bean extract is 0.3464558 times the odds in cucumber extract holding seeds type fixed.

(b) Residual deviance and pearson statistics show that the logistic regression model is poorly fixed with a p value smaller than 0.05, which suggests the potential issue of over dispersion. The half normal plot also generates a linear line deviated from the reference line. Estimated dispersion factor is 2.13 or 2.20 based on pearson residual or residual deviance. The updated model with constant dispersion factor gives a better fit of logistic model with p value larger than 0.05.

(c) Hierarchical sampling (clusters in the group has its own germination rate from the same distribution) might be the source of over dispersion as π becomes a random variable in this case.