P8131 Spring 2022 Homework #3 Solution

1. The following table gives data from a retrospective study between esophageal cancer and daily alcohol consumption adjusted for age.

Age	Case Daily Alcohol Consumption		Control Daily Alcohol Consumption	
	0-79 g	80+ g	0-79 g	80+ g
25-34	0	1	106	9
35-44	5	4	164	26
45-54	21	25	138	29
55-64	34	42	139	27
65-74	36	19	88	18
75+	8	5	31	0

   (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.

   (b) Let Ψ_j be the odds ratio relating alcohol consumption and disease in the jth age group ($j = 1, ..., 6$). Assume different age groups have different odds. Compare the following two models: $M_0 : \Psi_j = 1$ for all j;
   $M_1 : \Psi_j = \Psi$ (where Ψ is an unknown constant);
   (Hint: First write out the models and check if they are nested; if so, use deviance analysis to compare the two models.)

Solution:

   (a) We will use a logistic model to fit a prospective model. The outcome is disease, main predictor is exposure status, and covariate is age.

$$\log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 X + \beta_2 X_{age} = -5.02 + 1.78X + 0.06X_{age}$$

   $\exp(\beta_1)$ is the odds ratio for the association between disease and alcohol consumption, holding age fixed. With $\hat{\beta}_1 = 1.78$, the odds of having the disease among heavy drinkers is 5.93 (= exp(1.78)) times of that among non-heavy drinkers.

   $\exp(\beta_2)$ is the odds ratio for the association between disease and age, holding exposure status fixed. With $\hat{\beta}_2 = 0.06$, the odds of having the disease will increase by 6.2% (exp(0.06) = 1.062) for one level increase in age.

   (b) Outcome is disease status and main predictor is exposure status.

Model	Linear predictor	Deviance	df
M_0	α_j	90.56	6
M_1	$\alpha_j + \beta X$	11.04	5

Comparing the two models is equivalent to testing $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$. The test statistic is $90.56 - 11.04 = 79.52$, with p-value < 0.0001, based on a χ^2 distribution with $df = 6 - 5 = 1$. Null hypothesis is rejected and there is evidence that there is an association between disease and alcohol consumption.

2. The following table provides data from a study of the germination of two species of Orobanche seeds. The seeds were grown on 1/125 dilutions of two different root extract media (cucumber or bean) in a 2x2 factorial layout with replicates. The data (y_i/m_i) consist of the number of seeds, m_i, and the number germinating, y_i, for each batch. Interest focuses on the possible differences in germination rates for the two types of seed and root extract.

O. aegyptiaca 75		O. aegyptiaca 73	
Bean	Cucumber	Bean	Cucumber
10/39	5/6	8/16	3/12
23/62	53/74	10/30	22/41
23/81	55/72	8/28	15/30
26/51	32/51	23/45	32/51
17/39	46/79	0/4	3/7
	10/13		

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

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

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

Solution:

(a) Fit a model ignoring over-dispersion.

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

Where $X_{seed} = 1$ if O.aegyptiaca 75 seeds, 0 if O.aegyptiaca 73 seeds and $X_{root} = 1$ if bean extract, 0 if cucumber extract.

Parameter	Value
$\hat{\beta}_0$	0.36
$\hat{\beta}_1$	0.27
$\hat{\beta}_2$	-1.06
Pearson χ^2	38.31 (df=18)

$\exp(\hat{\beta}_0) = \exp(0.36) = 1.43$ is the odds of a O.aegyptiaca 73 seed grown in cucumber extract for germinating.

$\exp(\hat{\beta}_1)$ is the estimated odds ratio for comparing O.aegyptiaca 75 seeds and O.aegyptiaca 73 seeds, holding root extract fixed. With $\beta_1 = 0.27$, the odds of germinating in O.aegyptiaca 75 seeds is 1.31 ($= \exp(0.27)$) times of that in O.aegyptiaca 73 seeds, holding root extract fixed.

$\exp(\hat{\beta}_2)$ is the estimated odds ratio for comparing bean and cucumber extract amongst O.aegyptiaca 73 seeds, holding seed species fixed. With $\hat{\beta}_2 = -1.06$, the odds of germinating in bean extract is 0.35 ($= \exp(-1.06)$) times of that in cucumber extract, holding seed species fixed.

(b) Deviance and Pearson χ^2 statistics are large, suggesting poor fitting, which could be caused by ignoring over-dispersion. Half-normal plot using residuals from this model shows evidence of over-dispersion.

The dispersion parameter ϕ is 2.13 and $\tilde{\phi}$ is 2.20, both are greater than 1.

The updated model has the same coefficient estimates as those in part 1).

(c) A plausible cause of over-dispersion is the batch effect (i.e., different groups may have different germination rates which follow the same distribution). As a result, the response variables may follow a beta-binomial distribution instead of a binomial distribution.

Appendix

```
#################
### Problem 1 ###
#################
# 1
age <- rep(c(25, 35, 45, 55, 65, 75), 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)
exposure <- c(rep(1, 6), rep(0 , 6))
resp <- cbind(case, control)

logit.prosp=glm(resp~exposure+age, family=binomial(link='logit'))
summary(logit.prosp)

# 2
library(psych)
age_cat <- c(1:6)
age_cat <- factor(age_cat)
ind <- dummy.code(age_cat)
a1 <- rep(ind[,1], 2)
```

```
a2 <- rep(ind[,2], 2)
a3 <- rep(ind[,3], 2)
a4 <- rep(ind[,4], 2)
a5 <- rep(ind[,5], 2)
a6 <- rep(ind[,6], 2)

m0=glm(resp~a1+a2+a3+a4+a5+a6, family=binomial(link='logit'))
summary(m0)
sum(residuals(m0,type='deviance')^2)

m1=glm(resp~exposure+a1+a2+a3+a4+a5+a6, family=binomial(link='logit'))
summary(m1)
sum(residuals(m1,type='deviance')^2)

##################
### Problem 2 ###
##################
seed <- c(rep(1, 11), rep(0, 10)) # 1=O.aegyptiaca 75
root <- c(rep(1, 5), rep(0, 6), rep(1, 5), rep(0, 5)) # 1=bean
y <- c(10, 23, 23, 26, 17, 5, 53, 55, 32, 46, 10, 8, 10, 8, 23, 0,
       3, 22, 15, 32, 3)
m <- c(39, 62, 81, 51, 39, 6, 74, 72, 51, 79, 13, 16, 30, 28, 45,
       4, 12, 41, 30, 51, 7)

# 1
# fit binomial (logistic) without dispersion
none.disp=glm(cbind(y,m-y)~seed+root, family=binomial(link='logit'))
summary(none.disp)
1-pchisq(none.disp$deviance, 21-3)
sum(residuals(none.disp,type='pearson')^2)

# calc dispersion param
G.stat=sum(residuals(none.disp,type='pearson')^2) # pearson chisq
G.stat
phi=G.stat/(21-3)
phi
tilde.phi=none.disp$deviance/none.disp$df.residual
tilde.phi # similar to the one estimated from pearson chisq

########################
# 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='Ordered Abs Pearson Residuals')
abline(a=0,b=1)
abline(a=0,b=sqrt(phi),lty=2)

########################
# fit model with constant over-dispersion
summary(none.disp,dispersion=phi)
# goodness of fit
1-pchisq(none.disp$deviance/phi, 21-3)
# fit is ok
```