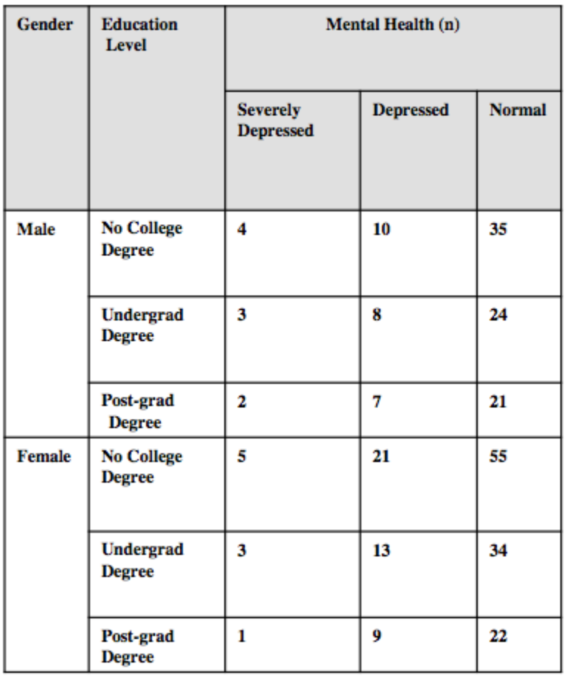
ADA Homework 8

Wenxin Liang UNI:wl2455

Consider the data in Table 1 on mental health.

Table 1



Gender	Education Level	Mental Health (n)		
		Severely Depressed	Depressed	Normal
Male	No College Degree	4	10	35
	Undergrad Degree	3	8	24
	Post-grad Degree	2	7	21
Female	No College Degree	5	21	55
	Undergrad Degree	3	13	34
	Post-grad Degree	1	9	22

Question 1

Categorize Mental Health as a binary variable, with values 0, if Normal, and 1, Otherwise; and Education Level with values 0 if No College Degree, and 1 otherwise.

Based on R, we categorized Mental Health and Education Level as followed,

```
> Data_Mental <- read.table("/Users/Wenxin_AN/Documents/Master/ada/MENTAL HEALTH.csv",sep=";",header=TRUE)
> Education <- 1*(!Data_Mental$Education.Level =="No College Degree")
> Mental <- 1*(!Data_Mental$Mental.Health == "Normal")
> table(Mental,Education)
         Education
Mental     0    1
      0   90 101
      1   40   46
```

a) Determine whether there is association between Education Level and Mental Health, using logistic regression, without adjusting for Gender. Interpret what the estimated parameters denote.

Based on the question since without adjusting for Gender, we have,

$$X_1 = \begin{cases} 0 & No\ College\ Degree \\ 1 & Otherwise \end{cases}$$

$$Y = \begin{cases} 0 & Normal \\ 1 & Otherwise \end{cases}$$

Let p_x denote the probability of having mental health (severely depressed, depressed) given education level of the individual. Then we have $p_x = \frac{e^{\beta_0+\beta_1 X_1}}{1+e^{\beta_0+\beta_1 X_1}}$

In terms of the link function we obtain,

$$logit(p_x) = \beta_0 + \beta_1 X_1$$

When $X_1 = 0$ which means when the education level is no college degree, $logit(p_x) = \beta_0$, which gives,

$$exp\{\beta_0\} = \frac{p_0}{1-p_0}$$

We obtain $exp\{\beta_0\}$ is the odds of having mental health (severely depressed, depressed) for no college degree.

When $X_1 = 1$ which means the education level is otherwise (undergrad degree or post-grad degree), $logit(p_x) = \beta_0 + \beta_1$, which gives,

$$\beta_1 = logit(p_1) - \beta_0 \;\; or \;\; \beta_1 = ln\frac{p_1}{1-p_1} - ln\frac{p_0}{1-p_0}$$

We obtain β_1 is the log odds ratio of having mental health (severely depressed, depressed) for the education level is otherwise (undergrad degree or post-grad degree).

Based on R, we get

```
> fit1_1 <- glm(Mental ~ Education,family = "binomial")
> summary(fit1_1)

Call:
glm(formula = Mental ~ Education, family = "binomial")

Deviance Residuals:
     Min          1Q     Median          3Q         Max
-0.8664   -0.8664   -0.8576     1.5243     1.5353

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.81093       0.19003    -4.267 1.98e-05 ***
Education       0.02445       0.26029     0.094       0.925
---
Signif. codes:   0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

     Null deviance: 343.19    on 276    degrees of freedom
Residual deviance: 343.18    on 275    degrees of freedom
AIC: 347.18
```

```
Number of Fisher Scoring iterations: 4
```

Therefore, we can conclude that,

$$\beta_0 = -0.81093$$

$$\beta_1 = 0.02445$$

Since $exp\{\beta_1\} = 1.024751 \approx 1$, all the coefficients of the estimated parameters are not significant so we conclude that there is no association between Education Level and Mental Health.

$\beta_0 = -0.81093$ denotes the log odds of having mental health (severely depressed, depressed) for no college degree.

$\beta_1 = 0.02445$ denotes the log odds ratio of having mental health (severely depressed, depressed) for the education level is otherwise (undergrad degree or post-grad degree) relative to education level is no college degree.

b) Repeat (a) adjusting for Gender. Interpret what the estimated parameters denote.

```
> Gender <- 1*(Data_Mental$Gender == "Male")
> table(Mental,Education,Gender)
, , Gender = 0

      Education
Mental  0  1
     0 55 56
     1 26 26

, , Gender = 1

      Education
Mental  0  1
     0 35 45
     1 14 20

> fit2 <- glm(Mental ~ Education+Gender, family ="binomial")
> summary(fit2)

Call:
glm(formula = Mental ~ Education + Gender, family = "binomial")

Deviance Residuals:
    Min        1Q    Median        3Q       Max
-0.8822   -0.8709   -0.8463    1.5047    1.5635

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept) -0.77389     0.21369   -3.622 0.000293 ***
Education    0.03093     0.26093    0.119 0.905633
Gender      -0.09946     0.26545   -0.375 0.707897
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 343.19  on 276  degrees of freedom
Residual deviance: 343.04  on 274  degrees of freedom
AIC: 349.04

Number of Fisher Scoring iterations: 4
```

Since $exp\{\beta_1\} = 1.031416 \approx 1$, all the coefficients of the estimated parameters are not significant so we conclude that there is no association between Education Level and Mental Health.
$\beta_0 = -0.77389$ denotes the log odds of having mental health (severely depressed, depressed) for no college degree and female.
$\beta_1 = 0.03093$ denotes the log odds ratio of having mental health (severely depressed, depressed) for the education level is otherwise (undergrad degree or post-grad degree) relative to no college degree and for any gender (male or female).
$\beta_2 = -0.09946$ denotes the log odds ratio of having mental health (severely depressed, depressed) for male relative to female for any education level (undergrad degree or post-grad degree).

```
> fit2_1 <- glm(Mental ~ Education+Gender+Education*Gender, family
="binomial")
> summary(fit2_1)

Call:
glm(formula = Mental ~ Education + Gender + Education * Gender,
    family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8799  -0.8733  -0.8576   1.5075   1.5829

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.74924    0.23800  -3.148  0.00164 **
Education        -0.01802    0.33610  -0.054  0.95725
Gender           -0.16705    0.39578  -0.422  0.67296
Education:Gender  0.12338    0.53403   0.231  0.81729
```

```
---
Signif. codes:   0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

     Null deviance: 343.19   on 276   degrees of freedom
Residual deviance: 342.99   on 273   degrees of freedom
AIC: 350.99

Number of Fisher Scoring iterations: 4
```

Since $exp\{\beta_1\} = 0.9821414 \approx 1$, all the coefficients of the estimated parameters are not significant so we conclude that there is no association between Education Level and Mental Health.

c) Assess whether it is appropriate to pool data across male and female subjects using a suitable logistic regression model.

Since all the coefficients of the estimated parameters in the model in part a which do not adjust for Gender and in the models in part b which adjusting for Gender are not significant then we can conclude are there is no association between Education Level and Mental Health and Gender do not have so much influence to Mental. Then it is appropriate to pool data across male and female subjects using a suitable logistic regression model.

Also we can use the woolf test to verify,

```
> woolf <- function(x) {
+    x <- x + 1 / 2
+    k <- dim(x)[3]
+    or <- apply(x, 3, function(x) (x[1,1]*x[2,2])/(x[1,2]*x[2,1]))
+    w <-   apply(x, 3, function(x) 1 / sum(1 / x))
+    1 - pchisq(sum(w * (log(or) - weighted.mean(log(or), w)) ^ 2), k - 1)
+ }
> woolf(table(Mental,Education,Gender))
[1] 0.826266
```

Since we obtain the p-value=0.826266>0.05 so we fail to reject the null hypothesis then we conclude that there is no significant heterogeneity. Then it is appropriate to pool data across male and female subjects.

Question 2

Repeat 1 (a) - 1 (c) above now using Educational Background as a trichotomous variable, i.e., No College Degree, Undergrad Degree, Post-grad Degree.

```
> # "No College Degree" as the reference group
> Edu.ud <- 1*(!Data_Mental$Education.Level =="Undergrad Degree")
> Edu.pd <- 1*(!Data_Mental$Education.Level =="Post-grad Degree")
>
```

Part a. Determine whether there is association between Education Level and Mental Health, using logistic regression, without adjusting for Gender. Interpret what the estimated parameters denote.

```
> # Part a
> fit3 <- glm(Mental ~ Edu.ud+Edu.pd, family = "binomial")
> summary(fit3)

Call:
glm(formula = Mental ~ Edu.ud + Edu.pd, family = "binomial")

Deviance Residuals:
     Min        1Q    Median       3Q      Max
-0.8743   -0.8576   -0.8576    1.5145    1.5380

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.770437    0.407771   -1.889    0.0588 .
Edu.ud      -0.046324    0.300648   -0.154    0.8775
Edu.pd       0.005831    0.334662    0.017    0.9861
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 343.19   on 276   degrees of freedom
Residual deviance: 343.16   on 274   degrees of freedom
AIC: 349.16

Number of Fisher Scoring iterations: 4
```

Since $exp\{\beta_1\} = 0.9547326 \approx 1$ and $exp\{\beta_2\} = 1.005848 \approx 1$, all the coefficients of the estimated parameters are not significant so we conclude that there is no association between Education Level and Mental Health.

$\beta_0 = -0.770437$ denotes the log odds of having mental health (severely depressed, depressed) relative to no college degree.

$\beta_1 = -0.046324$ denotes the log odds ratio of having mental health (severely depressed, depressed) for the education level is undergrad degree relative to education level is no college degree.

$\beta_2 = 0.005831$ denotes the log odds ratio of having mental health (severely depressed, depressed) for the education level is post-grad degree relative to education level is no college degree.

b) Repeat (a) adjusting for Gender. Interpret what the estimated parameters denote.

```
> fit4 <- glm(Mental ~ Edu.ud+Edu.pd+Gender, family = "binomial")
> summary(fit4)

Call:
glm(formula = Mental ~ Edu.ud + Edu.pd + Gender, family = "binomial")

Deviance Residuals:
    Min          1Q     Median          3Q         Max
-0.8889   -0.8707   -0.8374     1.5169     1.5630

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.720206      0.429767    -1.676      0.0938 .
Edu.ud          -0.049738      0.300866   -0.165      0.8687
Edu.pd          -0.004591      0.335937   -0.014      0.9891
Gender          -0.097709      0.265819   -0.368      0.7132
---
Signif. codes:   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 343.19    on 276    degrees of freedom
Residual deviance: 343.02    on 273    degrees of freedom
AIC: 351.02

Number of Fisher Scoring iterations: 4
```

Since $exp\{\beta_1\} = 0.9514787 \approx 1$ and $exp\{\beta_2\} = 0.9954195 \approx 1$, all the coefficients of the estimated parameters are not significant so we conclude that there is no association between Education Level and Mental Health when adjusting for Gender.

$\beta_0 = -0.720206$ denotes the log odds of having mental health (severely depressed, depressed) for no college degree and female.

$\beta_1 = -0.049738$ denotes the log odds ratio of having mental health (severely depressed, depressed) for the education level undergrad degree relative to no college degree and for any gender (male or female).

$\beta_2 = -0.004591$ denotes the log odds ratio of having mental health (severely depressed, depressed) for the education level post-grad degree relative to no college degree and for any gender (male or female).

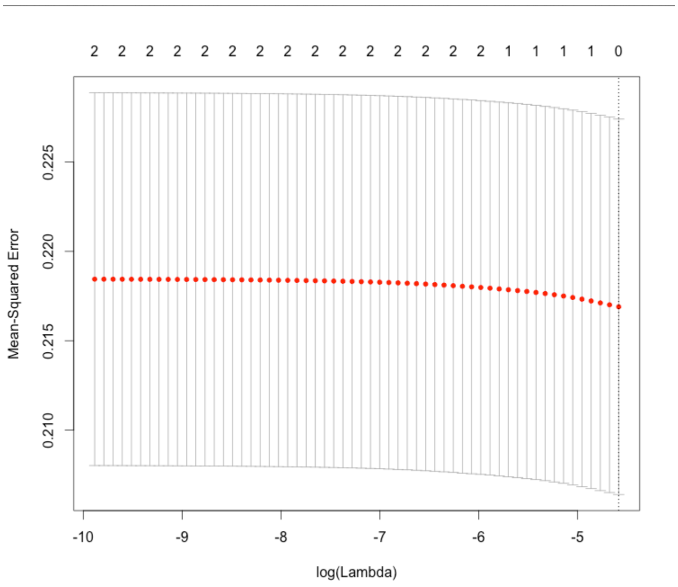
$\beta_3 = -0.097709$ denotes the log odds ratio of having mental health (severely depressed, depressed) for male relative to female for any education level (undergrad degree or post-grad degree).

c) Assess whether it is appropriate to pool data across male and female subjects using a suitable logistic regression model.

Since all the coefficients of the estimated parameters in the model in part a which do not adjust for Gender and in the models in part b which adjusting for Gender are not significant then we can conclude are there is no association between Education Level and Mental Health and Gender do not have so much influence to Mental. Then it is appropriate to pool data across male and female subjects using a suitable logistic regression model. Also here we can conclude that it is appropriate to pool data across Education level subjects using a suitable logistic regression model.

Question 3 Repeat 1 (b) using the lasso

```
> library(glmnet)
> X <- model.matrix(Mental~Education+Gender)
> y <- Mental
> fit <- glmnet(X,y)
> cvfit <- cv.glmnet(X,y)
> plot(cvfit)
> cv_out <- cv.glmnet(X,y,alpha=1,family="binomial")
> bestlammin <- cv_out$lambda.min
> result <- glmnet(X,y,alpha=1,family="binomial")
> lasso.coef <- predict(result,type="coefficients",s=bestlammin)
> lasso.coef
4 x 1 sparse Matrix of class "dgCMatrix"
                          1
(Intercept) -0.7979261
(Intercept)    .
Education      .
Gender         .
```



Based Lasso procedure the final model includes only the intercept term, which means we drop down all the predictors, Education Level and Gender, since the other predictors are not significant. This result is corresponding to the result we obtain in Question 1 part b. Then we conclude that there is no association between Education Level and Mental Health.

The code followed,

```
#Homework 8
# Question 1
Data_Mental <- read.table("/Users/Wenxin_AN/Documents/Master/ada/MENTAL
HEALTH.csv",sep=";",header=TRUE)
Education <- 1*(!Data_Mental$Education.Level =="No College Degree")
Mental <- 1*(!Data_Mental$Mental.Health == "Normal")
table(Mental,Education)

# Part a
fit1_1 <- glm(Mental ~ Education,family = "binomial")
summary(fit1_1)
exp(0.02445)

# Part b
Gender <- 1*(Data_Mental$Gender == "Male")
table(Mental,Education,Gender)
fit2 <- glm(Mental ~ Education+Gender, family ="binomial")
summary(fit2)
exp(0.030933)
fit2_1 <- glm(Mental ~ Education+Gender+Education*Gender, family ="binomial")
summary(fit2_1)
exp(-0.01802)
# Part c
woolf <- function(x) {
   x <- x + 1 / 2
   k <- dim(x)[3]
   or <- apply(x, 3, function(x) (x[1,1]*x[2,2])/(x[1,2]*x[2,1]))
   w <-    apply(x, 3, function(x) 1 / sum(1 / x))
   1 - pchisq(sum(w * (log(or) - weighted.mean(log(or), w)) ^ 2), k - 1)
}
woolf(table(Mental,Education,Gender))

# Question 2
# "No College Degree" as the reference group
Edu.ud <- 1*(!Data_Mental$Education.Level =="Undergrad Degree")
Edu.pd <- 1*(!Data_Mental$Education.Level =="Post-grad Degree")
```

```
# Part a
fit3 <- glm(Mental ~ Edu.ud+Edu.pd, family = "binomial")
summary(fit3)
exp(-0.046324)
exp(0.005831)

# Part b
fit4 <- glm(Mental ~ Edu.ud+Edu.pd+Gender, family = "binomial")
summary(fit4)
exp(-0.049738)
exp(-0.004591)

# Part c


# Question 3
library(glmnet)
X <- model.matrix(Mental~Education+Gender)
y <- Mental
fit <- glmnet(X,y)
cvfit <- cv.glmnet(X,y)
plot(cvfit)
cv_out <- cv.glmnet(X,y,alpha=1)
bestlammin <- cv_out$lambda.min
result <- glmnet(X,y,alpha=1)
lasso.coef <- predict(result,type="coefficients",s=bestlammin)
lasso.coef
```