Homework 7

Qianbo Wang
uni: qw2180

Consider the ChickWeight data in R. The body weights of the chicks were measured at birth (i.e., time=0) and every second day thereafter until day 20. They were also measured on day 21. There were four groups of chicks on different protein diets.

Categorize weight as a binary variable, with Weight Group = 1 (or Low), if weight > 215 mg, and 0, Otherwise.

Problem 1

Consider comparing Diet Levels 1 and 4 on Day 21.

(a). Determine whether there is association between Diet and Weight, using logistic regression, without adjusting for Birth Weight. Interpret what the estimated parameters denote.

Construct a logistic regression model with adjusting for Birth Weight with Weight for Diet group 1 and 4 on day 21. Restructure the data as categorical data type. The response and explanatory variable is like:

$$\text{Weight}_i = \begin{cases} 1 & \text{Weight} > 215 \\ 0 & \text{Weight} \leq 215 \end{cases} \qquad \text{Group}_i = \begin{cases} 1 & \text{Diet} = 1 \\ 0 & \text{Diet} \neq 1 \end{cases}$$

Then the result is as follows:

Logistic Regression without adjust result

```
Call:
glm(formula = Weight ~ Group, family = "binomial", data = sub_day21)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2735  -0.6444  -0.6444   1.0842   1.8297

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.2231     0.6708   0.333   0.7394
Group        -1.6895     0.9275  -1.822   0.0685 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 31.343  on 24  degrees of freedom
Residual deviance: 27.808  on 23  degrees of freedom
AIC: 31.808

Number of Fisher Scoring iterations: 4
```

The model is:

$$\text{logit(p)} = 0.2231 - 1.6895 * \text{Group}$$

So for Diet group 1, the model is:

$$\text{logit(p)} = 0.2231 - 1.6895$$

And for Diet group 4, the model is:

$$\text{logit(p)} = 0.2231$$

There are two parameters here, intercept β_0, and coefficient β_1 on Diet group.

- β_0 denotes the log odds ratio of Weight > 215 for Diet group 4, i.e., the odds ratio of Weight > 215 for Diet group 4 is $e^0.2231$.
- β_1 denotes the log odds ratio of Weight > 215 for Diet group 1 relative to Diet Group 4, i.e., odds ratio of Weight > 215 for Diet group 1 relative to Diet Group 4 is $e^{-1.6895}$, and also the odds ratio of Weight > 215 for Diet group 1 is $e^{-1.4664}$.

Since p-value for the intercept and Group are both p-value > 0.05, so we don't reject the null, i.e., there are no significant association between Diet Group 1 and 4 and Categorical Weight without adjusting for BirthWeight.

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

Construct a logistic regression model without adjusting for Birth Weight with Weight for Diet group 1 and 4 on day 21. Restructure the data as categorical data type. The response and explanatory variable is like:

$$\text{Weight}_i = \begin{cases} 1 & \text{Weight} > 215 \\ 0 & \text{Weight} \leq 215 \end{cases} \qquad \text{Group}_i = \begin{cases} 1 & \text{Diet} = 1 \\ 0 & \text{Diet} = 4 \end{cases}$$

Then the result is as follows:

Logistic Regression with adjust result

```
Call:
glm(formula = Weight ~ Group + BirthWeight, family = "binomial",
    data = sub_day21)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3159  -0.7680  -0.4050   0.6028   1.6956

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  57.2079    31.1175   1.838   0.0660 .
Group        -1.2935     1.0467  -1.236   0.2165
BirthWeight  -1.3899     0.7567  -1.837   0.0662 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 31.343  on 24  degrees of freedom
Residual deviance: 22.856  on 22  degrees of freedom
AIC: 28.856
```

```
Number of Fisher Scoring iterations: 5
```

The model is:

$$\text{logit(p)} = 57.2079 - 1.2935 * \text{Group} - 1.3899 * \text{BirthWeight}$$

So for Diet group 1 the model is:

$$\text{logit(p)} = 57.2079 - 1.2935 - 1.3899 * \text{BirthWeight}$$

And for Diet group 4 the model is:

$$\text{logit(p)} = 57.2079 - 1.3899 * \text{BirthWeight}$$

There are three parameters here, intercept β_0, and coefficient β_1 on Diet group, and coefficient β_2 on BirthWeight.

- β_0 denotes when BirthWeight is given 0, the log odds ratio of Weight > 215 for Diet group 4, i.e., when BirthWeight=0(which is not realistic), the odds ratio of Weight > 215 for Diet group 4 is $e^{57.2079}$, and the odds ratio of Weight > 215 for Diet group 4 when given BirthWeight $= x$ is $e^{57.2079-1.3899x}$.
- β_1 denotes the log odds ratio of Weight > 215 for Diet group 1 relative to Diet Group 4 when given BirthWeight $= 0$(which is not realistic), i.e., when BirthWeight $= 0$ odds ratio of Weight > 215 for Diet group 1 relative to Diet Group 4 is $e^{-1.3899}$, and odds ratio of Weight > 215 for Diet group 1 when given BirthWeight $= x$ is $e^{55.9144-1.3899x}$
- β_2 denotes under same Diet Group, the change in log odds for Weight > 215 when the BirthWeight is different, i.e., under same Diet Group, 1 unit change in BirthWeight will cause the odds for Weight > 215 in day 21 change $e^{-1.3899}$.

Since p-value for the intercept, Group and BirthWeight are all p-value > 0.05, so we don't reject the null, i.e., there are no significant association between Diet Group 1 and 4 and Categorical Weight with adjusting for BirthWeight.

Problem 2

Repeat 1 for all 4 Diet Levels.

(a). Without adjusting for BirthWeight.

Construct a logistic regression model without adjusting for Birth Weight with Weight for Diet group 1 and 4 on day 21. Restructure the data as categorical data type. The response and explanatory variable is like:

$$\text{Weight}_i = \begin{cases} 1 & \text{Weight} > 215 \\ 0 & \text{Weight} \leq 215 \end{cases} \qquad \text{Group2}_i = \begin{cases} 1 & \text{Diet} = 2 \\ 0 & \text{Diet} \neq 2 \end{cases}$$

$$\text{Group1}_i = \begin{cases} 1 & \text{Diet} = 1 \\ 0 & \text{Diet} \neq 1 \end{cases} \qquad \text{Group3}_i = \begin{cases} 1 & \text{Diet} = 3 \\ 0 & \text{Diet} \neq 3 \end{cases}$$

Then the result is as follows:

Logistic Regression without adjust result

```
Call:
glm(formula = Weight ~ Group1 + Group2 + Group3, family = "binomial",
    data = sub_day21_all)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7941  -0.6444  -0.6444   1.0842   1.8297

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.2231     0.6708   0.333   0.7394
Group1       -1.6895     0.9275  -1.822   0.0685 .
Group2       -0.2231     0.9220  -0.242   0.8088
Group3        1.1632     1.0368   1.122   0.2619
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 62.183  on 44  degrees of freedom
Residual deviance: 51.679  on 41  degrees of freedom
AIC: 59.679

Number of Fisher Scoring iterations: 4
```

The model is:

$$\text{logit(p)} = 0.2231 - 1.6895 * \text{Group1} - 0.2231 * \text{Group2} + 1.1632 * \text{Group3}$$

So for Diet group 1, the model is:

$$\text{logit(p)} = 0.2231 - 1.6895 = -1.4664$$

And for Diet group 2, the model is:

$$\text{logit(p)} = 0.2231 - 0.2231 = 0$$

And for Diet group 3, the model is:

$$\text{logit(p)} = 0.2231 + 1.1632 = 1.3863$$

And for Diet group 4, the model is:

$$\text{logit(p)} = 0.2231$$

There are four parameters here, intercept β_0, and coefficient β_1 on Diet group1, coefficient β_2 on Diet group2, coefficient β_3 on Diet group3.

- β_0 denotes the log odds ratio of Weight > 215 for Diet group 4, i.e., the odds ratio of Weight > 215 for Diet group 4 is $e^{0.2231}$.
- β_1 denotes the log odds ratio of Weight > 215 for Diet group 1 relative to Diet Group 4, i.e., odds ratio of Weight > 215 for Diet group 1 relative to Diet Group 4 is $e^{-1.6895}$, and odds ratio of Weight > 215 for Diet group 1 is $e^{-1.4664}$.
- β_2 denotes the log odds ratio of Weight > 215 for Diet group 2 relative to Diet Group 4, i.e., odds ratio of Weight > 215 for Diet group 2 relative to Diet Group 4 is $e^{-0.2231}$, and odds ratio of Weight > 215 for Diet group 2 is $e^0 = 1$.

- β_3 denotes the log odds ratio of Weight > 215 for Diet group 3 relative to Diet Group 4, i.e., odds ratio of Weight > 215 for Diet group 3 relative to Diet Group 4 is $e^{1.1632}$, and odds ratio of Weight > 215 for Diet group 3 is $e^{1.3863}$.

Since p-value for the intercept, Group1, Group2, Group3 are all p-value > 0.05, so we don't reject the null, i.e., there are no significant association between Diet Group from 1 to 4 and Categorical Weight without adjusting for BirthWeight.

(b). With adjusting for BirthWeight.

Construct a logistic regression model without adjusting for Birth Weight with Weight for Diet group 1 and 4 on day 21. Restructure the data as categorical data type. The response and explanatory variable is like:

$$\text{Weight}_i = \begin{cases} 1 & \text{Weight} > 215 \\ 0 & \text{Weight} \le 215 \end{cases} \qquad \text{Group2}_i = \begin{cases} 1 & \text{Diet} = 2 \\ 0 & \text{Diet} \ne 2 \end{cases}$$

$$\text{Group1}_i = \begin{cases} 1 & \text{Diet} = 1 \\ 0 & \text{Diet} \ne 1 \end{cases} \qquad \text{Group3}_i = \begin{cases} 1 & \text{Diet} = 3 \\ 0 & \text{Diet} \ne 3 \end{cases}$$

Then the result is as follows:

Logistic Regression with adjust result

```
Call:
glm(formula = Weight ~ Group1 + Group2 + Group3 + BirthWeight,
    family = "binomial", data = sub_day21_all)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3302  -0.7262  -0.4018   0.8540   1.7100

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  26.3751    14.6987   1.794   0.0728 .
Group1       -1.3805     0.9641  -1.432   0.1522
Group2       -0.3800     0.9932  -0.383   0.7020
Group3        1.1864     1.0707   1.108   0.2678
BirthWeight  -0.6389     0.3582  -1.784   0.0745 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 62.183  on 44  degrees of freedom
Residual deviance: 48.021  on 40  degrees of freedom
AIC: 58.021

Number of Fisher Scoring iterations: 4
```

The model is:

$$\text{logit(p)} = 26.3751 - 1.3805 * \text{Group1} - 0.3800 * \text{Group2} + 1.1864 * \text{Group3} - 0.6389 * \text{BirthWeight}$$

So for Diet group 1 the model is:

$$\text{logit(p)} = 26.3751 - 1.3805 - 0.6389 * \text{BirthWeight}$$

And for Diet group 2 the model is:

$$\text{logit(p)} = 26.3751 - 0.3800 - 0.6389 * \text{BirthWeight}$$

And for Diet group 3 the model is:

$$\text{logit(p)} = 26.3751 + 1.1864 - 0.6389 * \text{BirthWeight}$$

And for Diet group 4 the model is:

$$\text{logit(p)} = 26.3751 - 0.6389 * \text{BirthWeight}$$

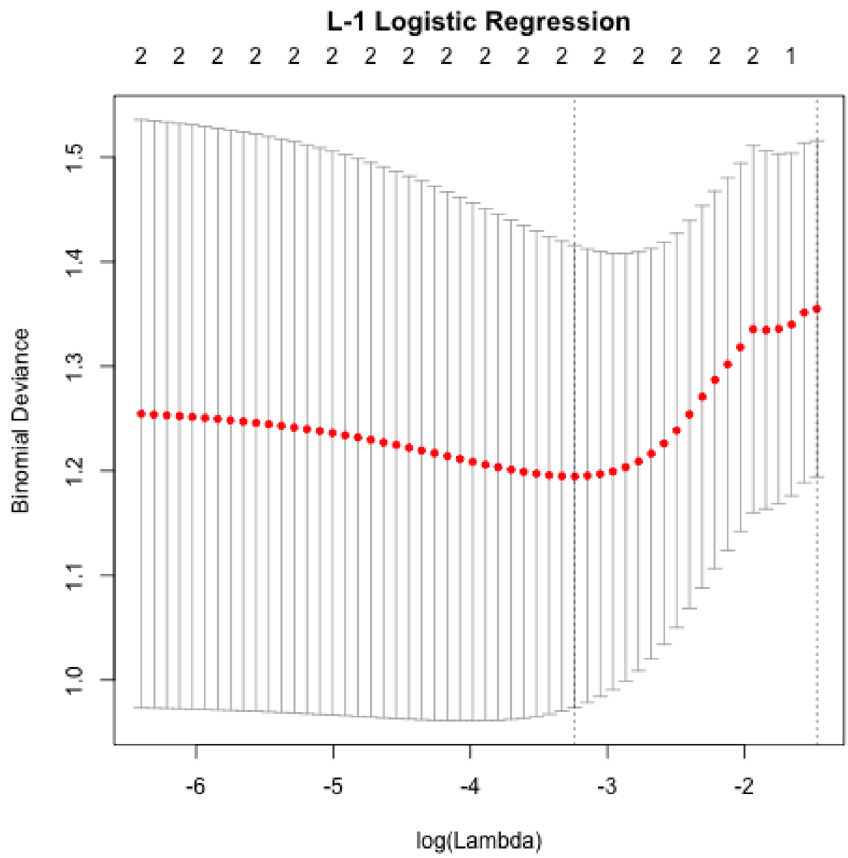
There are five parameters here, intercept β_0, and coefficient β_1 on Diet group1, coefficient β_2 on Diet group2, coefficient β_3 on Diet group3, and coefficient β_4 on BirthWeight.

- β_0 denotes the log odds ratio of Weight > 215 for Diet group 4 under given BirthWeight=0(which is not realistic), i.e., the odds ratio of Weight > 215 for Diet group 4 when BirthWeight=0 is $e^{26.3751}$, and the odds ratio of Weight > 215 for Diet group 4 under given BirthWeight $= x$ is $e^{26.3751-0.6389x}$.
- β_1 denotes the log odds ratio of Weight > 215 for Diet group 1 relative to Diet Group 4, under given BirthWeight=0(which is not realistic), i.e., when BirthWeight=0, odds ratio of Weight > 215 for Diet group 1 relative to Diet Group 4 is $e^{-1.3805}$, and under given BirthWeight $= x$ odds ratio of Weight > 215 for Diet group 1 relative to Diet Group 4 is $e^{-1.3805-0.6389x}$.
- β_2 denotes the log odds ratio of Weight > 215 for Diet group 2 relative to Diet Group 4, under given BirthWeight=0(which is not realistic), i.e., when BirthWeight=0, odds ratio of Weight > 215 for Diet group 2 relative to Diet Group 4 is $e^{1.1864}$, and under given BirthWeight $= x$ odds ratio of Weight > 215 for Diet group 2 relative to Diet Group 4 is $e^{1.1864-0.6389x}$.
- β_3 denotes the log odds ratio of Weight > 215 for Diet group 3 relative to Diet Group 4, under given BirthWeight=0(which is not realistic), i.e., when BirthWeight=0, odds ratio of Weight > 215 for Diet group 2 relative to Diet Group 4 is $e^{1.1864}$, and under given BirthWeight $= x$ odds ratio of Weight > 215 for Diet group 3 relative to Diet Group 4 is $e^{1.1864-0.6389x}$.
- β_4 denotes under same Diet Group, the change in log odds for Weight > 215 when the BirthWeight is changing, i.e., under same Diet Group, 1 unit change in BirthWeight will cause the odds for Weight > 215 in day 21 change $e^{-0.6389}$.

Since p-value for the intercept, Group1, Group2, Group3 and BirthWeight are all p-value > 0.05, so we don't reject the null, i.e., there are no significant association between Diet Group from 1 to 4 and Categorical Weight with adjusting for BirthWeight.

Problem 3

Repeat 1 using the L-1 regularized logistic regression.
We should use BirthWeight and Group for L-1 logistic regression. The cross validation plot of choosing best gamma is as follows:



Then, choose the best lambda, $lambda = 0.03921473$, and get the model with the best lambda. The coefficients are as follows:

Coefficients of L-1 Logistic Regression

```
3 x 1 sparse Matrix of class "dgCMatrix"
                     1
(Intercept) 39.9300576
BirthWeight -0.9754835
Group       -0.8577307
```

So, the final model is:

$$\text{logit(p)} = 39.9300576 - 0.9754835 * \text{BirthWeight} - 0.8577307 * \text{Group}$$

R Code:

```
rm(list=ls())
data("ChickWeight")

#Categorize weight as Weight
ChickWeight$Weight<-rep(0,dim(ChickWeight)[1])
ChickWeight$Weight[ChickWeight$weight>215]<-1

#logistic regression with Diet and Weight on day 21 without adjust for Birth Weight
sub_day21<-subset(ChickWeight,Diet %in% c(1,4) & Time==21)
sub_day21$Group<-rep(0,dim(sub_day21)[1])
sub_day21$Group[sub_day21$Diet==1]<-1
log_day21<-glm(Weight~Group,family="binomial",data=sub_day21)

sink('/Users/raymond/Drive/STAT W4201/HW7/log_day21.txt')
summary(log_day21)
sink()

BirthWeight<-subset(ChickWeight,select=c(weight,Chick),Time==0)
names(BirthWeight)[names(BirthWeight)=="weight"]<-"BirthWeight"
sub_day21<-merge(sub_day21,BirthWeight,by.y = "Chick")

#logistic regression with Diet and Weight on day 21 with adjust for Birth Weight
log_day21_adjust<-glm(Weight~Group+BirthWeight,family="binomial",data=sub_day21)

sink('/Users/raymond/Drive/STAT W4201/HW7/log_day21_adjust.txt')
summary(log_day21_adjust)
sink()

#all groups
sub_day21_all<-subset(ChickWeight,Time==21)
sub_day21_all$Group1<-rep(0,dim(sub_day21_all)[1])
sub_day21_all$Group1[sub_day21_all$Diet==1]<-1
sub_day21_all$Group2<-rep(0,dim(sub_day21_all)[1])
sub_day21_all$Group2[sub_day21_all$Diet==2]<-1
sub_day21_all$Group3<-rep(0,dim(sub_day21_all)[1])
sub_day21_all$Group3[sub_day21_all$Diet==3]<-1
log_day21_all<-glm(Weight~Group1+Group2+Group3,family="binomial",data=sub_day21_all)
sink('/Users/raymond/Drive/STAT W4201/HW7/log_day21_all.txt')
summary(log_day21_all)
sink()

#logistic regression with Diet and Weight on day 21 with adjust for Birth Weight
sub_day21_all<-merge(sub_day21_all,BirthWeight,by.y = "Chick")
log_day21_all_adjust<-glm(Weight~Group1+Group2+Group3+BirthWeight,family="binomial",data=sub_day21_all)
sink('/Users/raymond/Drive/STAT W4201/HW7/log_day21_all_adjust.txt')
summary(log_day21_all_adjust)
sink()

#L-1 regularized logistic regression with Diet and Weight on day 21 with adjust for Birth Weight
library(glmnet)
data<-as.matrix(subset(sub_day21,select = c("BirthWeight","Group")))
colnames(data)<-c("BirthWeight","Group")
l1log_day21<-glmnet(data,sub_day21$Weight,family="binomial")
cv_l1log_day21<-cv.glmnet(data,sub_day21$Weight,family="binomial")

png(filename = "/Users/raymond/Drive/STAT W4201/HW7/cv_l1log.png")
plot(cv_l1log_day21)
title(main = "L-1 Logistic Regression", line = 2.5)
dev.off()

lambda<-cv_l1log_day21$lambda.min
model<-cv_l1log_day21$glmnet.fit

coeff<-coef(model,lambda)
sink('/Users/raymond/Drive/STAT W4201/HW7/coeff.txt')
coeff
```

```
sink()
```