Qiuying Li

UNI: ql2280

HW6

Problem 1

Consider comparing Diet Levels 1 and 2 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.

Logistic Regression without adjust result

```
##Call:
##glm(formula = Weight ~ Group, family = "binomial", data = sub_day21)
## Deviance Residuals:
##      Min        1Q    Median       3Q       Max
## -1.48230  -1.01077  -0.05513   1.01382   1.35373
##
## Coefficients:
##            Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.4055     0.6455  -0.628    0.530
## Group        1.0986     1.0801   1.017    0.309
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22.181  on 15  degrees oxf freedom
## Residual deviance: 21.098 on 14  degrees of freedom
## AIC: 25.098
##
## Number of Fisher Scoring iterations: 4
```

The model is: $\text{logit}(p) = -0.4055 + 1.0986 * \text{Group}$

For Diet group 1, the model is:

$\text{logit}(p) = -0.4055 + 1.0986 = 0.6931$

For Diet group 4, the model is:

logit(p) = -0.4055

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

- β_0 denotes the log odds ratio of Weight <180 for Diet group 4, i.e., the odds ratio of Weight <180 for Diet group 4 is $e^{-0.4055.}$

- β_1 denotes the log odds ratio of Weight <180 for Diet group 1 relative to Diet Group 4, i.e., odds ratio of Weight <180 for Diet group 1 relative to Diet Group 4 is $e^{1.0986}$, and also the odds ratio of Weight <180 for Diet group 1 is $e^{0.6931}$.

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 2 on day 21. Restructure the data as categorical data type. The response and explanatory variable is like:

Logistic Regression with adjust result

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

```
##
## Deviance Residuals:
##      Min        1Q    Median       3Q       Max
## -1.83779  -0.41249   0.01454   0.63916   1.38171
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -80.539     42.885  -1.878   0.0604 .
## Group          0.597      1.431   0.417    0.6765
## BirthWeight    1.953      1.038   1.882   0.0598 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 22.181  on 15  degrees of freedom
## Residual deviance: 12.414  on 13  degrees of freedom
## AIC: 18.414
##
## Number of Fisher Scoring iterations: 6
```

The model is:

logit(p) = -80.539 – 0.597 * Group – 1.953 * BirthWeight

So for Diet group 1 the model is:

logit(p) = -80.539+0.597+1.953*BirthWeight = -79.942+1.953*BirthWeight

for Diet group 2 the model is:

logit(p) = -80.539+1.953*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 <180 for Diet group 2, i.e., when BirthWeight=0(which is not realistic), the odds ratio of Weight <180 for Diet group 2 is $e^{-80.539}$, and the odds ratio of Weight <180 for Diet group 4 when given BirthWeight = x is $e^{-80.539 - 0.597 * \text{ Group}}$.

- β_1 denotes the log odds ratio of Weight <180 for Diet group 1 relative to Diet Group 2 when given BirthWeight = 0(which is not realistic), i.e., when BirthWeight = 0 odds ratio of Weight <180 for Diet group 1 relative to Diet Group 2 is $e^{1.953}$, and odds ratio of Weight <180 for Diet group 1 when given BirthWeight is $e^{-79.942+1.953*\text{BirthWeight}}$

- β_2 denotes under same Diet Group, the change in log odds for Weight <180 when the BirthWeight is different, i.e., under same Diet Group, 1 unit change in BirthWeight will cause the odds for Weight <180 in day 21 change $e^{1.953}$.

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 2 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:

Repeat 1 for all 4 Diet Levels.

2. (a). Without adjusting for BirthWeight.

Logistic Regression without adjust result
Call:
glm(formula = biweight ~ Group1 + Group2 + Group3, family = "binomial",
data = sub.all)
Deviance Residuals:
Min 1Q Median 3Q Max
-1.28583 -1.01077 -0.00013 1.07272 1.79412
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -18.57 2174.21 -0.009 0.993
Group1 18.82 2174.21 0.009 0.993
Group2 18.16 2174.21 0.008 0.993
Group3 17.18 2174.21 0.008 0.994
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 57.286 on 44 degrees of freedom
Residual deviance: 45.398 on 41 degrees of freedom
AIC: 53.398
##
Number of Fisher Scoring iterations: 17

The model is:

logit(p) = -18.57+18.82*Group1+18.16*Group2+17.18*Group3

For Diet1, logit(p)=-18.57+18.82=0.25

For Diet2, logit(p)=-18.57+18.16=-0.41

For Diet3, logit(p)=-18.57+17.18=-1.39

For Diet4, logit(p)=-18.57

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 <180 for Diet group 4, i.e., the odds ratio of Weight <180 for Diet group 4 is $e^{-18.57}$.

• β_1 denotes the log odds ratio of Weight <180 for Diet group 1 relative to Diet Group 4, i.e., odds ratio of Weight <180 for Diet group 1 relative to Diet Group 4 is $e^{18.82}$, and odds ratio of Weight <180 for Diet group 1 is $e^{0.25}$.

• β_2 denotes the log odds ratio of Weight <180 for Diet group 2 relative to Diet Group 4, i.e., odds ratio of Weight <180 for Diet group 2 relative to Diet Group 4 is $e^{18.16}$, and odds ratio of Weight <180 for Diet group 2 is $e^{0.41}$.

• β_3 denotes the log odds ratio of Weight <180 for Diet group 3 relative to Diet Group 4, i.e., odds ratio of Weight <180 for Diet group 3 relative to Diet Group 4 is $e^{17.18}$, and odds ratio of Weight <180 for Diet group 3 is $e^{-1.39}$.

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.

2. (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:

Logistic Regression with adjust result
Call:
glm(formula = biweight ~ Group1 + Group2 + Group3 + BirthWeight,
family = "binomial", data = sub.all.birth)
Deviance Residuals:
Min 1Q Median 3Q Max
-1.72563 -0.80750 -0.00018 0.94758 2.35262
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -45.8699 2079.4814 -0.022 0.9824
Group1 18.4977 2079.4211 0.009 0.9929
Group2 18.3069 2079.4211 0.009 0.9930
Group3 17.2229 2079.4212 0.008 0.9934
BirthWeight 0.6652 0.3827 1.738 0.0822 .
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 57.286 on 44 degrees of freedom
Residual deviance: 41.904 on 40 degrees of freedom
AIC: 51.904
Number of Fisher Scoring iterations: 17

the model

logit(p) = -45.8699+18.49772*Group1+18.3069*Group2+17.2229*Group3+0.6652*birthweight

For Diet1, logit(p)=-45.8699+18.49772+0.6652*birthweight

For Diet2, logit(p)=-45.8699+18.3069+0.6652*birthweight

For Diet3, logit(p)=-45.8699+17.2229+0.6652*birthweight

For Diet4, logit(p)=-45.8699+0.6652*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 < 180 for Diet group 4 under given BirthWeight=0(which is not realistic), i.e., the odds ratio of Weight < 180 for Diet group 4 when BirthWeight=0 is $e^{-45.8699}$, and the odds ratio of Weight < 180 for Diet group 4 under given BirthWeight is $e^{-45.8699 + 0.6652*birthweight}$.

- β_1 denotes when BirthWeight=0, odds ratio of Weight < 180 for Diet group 1 relative to Diet Group 4 is $e^{-27.37218}$, and under given BirthWeight = x odds ratio of Weight < 180 for Diet group 1 relative to Diet Group 4 is $e^{-27.37218 + 0.6652*birthweight}$.

- β_2 denotes when BirthWeight=0, odds ratio of Weight < 180 for Diet group 2 relative to Diet Group 4 is $e^{-27.563}$, and under given BirthWeight = x odds ratio of Weight < 180 for Diet group 2 relative to Diet Group 4 is $e^{-27.5638 + 0.6652*birthweight}$.

- β_3 denotes when BirthWeight=0, odds ratio of Weight < 180 for Diet group 2 relative to Diet Group 4 is $e^{-28.647}$, and under given BirthWeight = x odds ratio of Weight < 180 for Diet group 3 relative to Diet Group 4 is $e^{-28.647 + 0.6652*birthweight}$

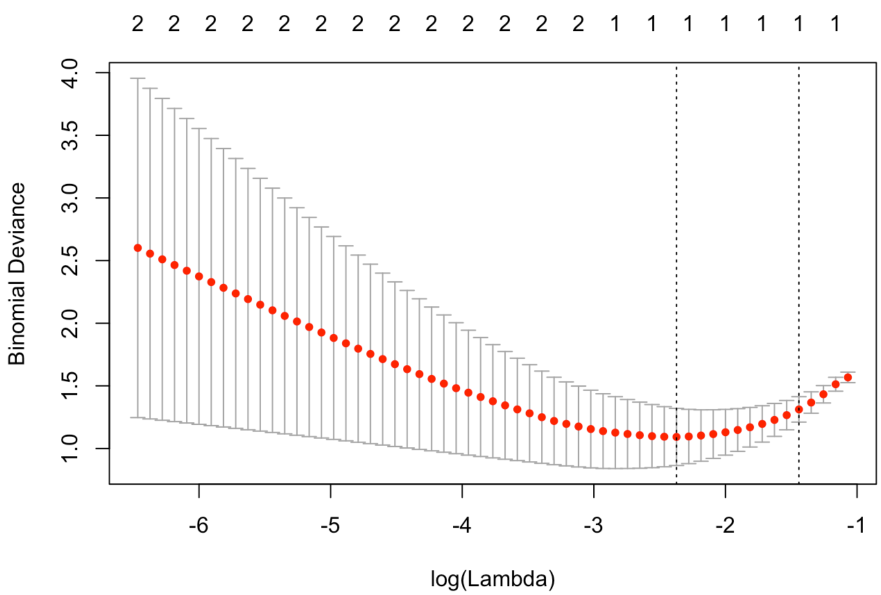
- β_4 denotes under same Diet Group, the change in log odds for Weight < 180 when the BirthWeight is changing, i.e., under same Diet Group, 1 unit change in BirthWeight will cause the odds for Weight < 180 in day 21 change $e^{0.6652}$.

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 Birth Weight 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) -39.88938   ## BirthWeight 0.97229   ## Group .
So, the final model is: logit(p) = -39.88938 + 0.97229 * BirthWeigh