Glm1

2023-10-21

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 WeightGroup = 1 (or Low), if weight < 110 mg, and 0, Otherwise.

Q1(a). Consider comparing Diet Levels 1 and 4 on Day 21.

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

```
# import data
library(datasets)
data("ChickWeight")
summary(ChickWeight)
```

```
##      weight           Time           Chick     Diet
##  Min.   : 35.0   Min.   : 0.00   13     : 12   1:220
##  1st Qu.: 63.0   1st Qu.: 4.00   9      : 12   2:120
##  Median :103.0   Median :10.00   20     : 12   3:120
##  Mean   :121.8   Mean   :10.72   10     : 12   4:118
##  3rd Qu.:163.8   3rd Qu.:16.00   17     : 12
##  Max.   :373.0   Max.   :21.00   19     : 12
##                                  (Other):506
```

```
# get baseline
birth_weight <- ChickWeight[ChickWeight$Time == 0, c("Chick", "weight")]
colnames(birth_weight) <- c("Chick", "weight_initial")
chickWeight <- merge(ChickWeight, birth_weight, by = "Chick", all.x = TRUE)

# get WeightGroup
chickWeight$WeightGroup <- ifelse(chickWeight$weight < 110, 1, 0)
chickWeight$Diet4 <- ifelse(chickWeight$Diet == 4, 1, 0)
chickWeight$Diet1 <- ifelse(chickWeight$Diet == 1, 1, 0)
chickWeight$Diet2 <- ifelse(chickWeight$Diet == 2, 1, 0)
chickWeight$Diet3 <- ifelse(chickWeight$Diet == 3, 1, 0)

summary(chickWeight)
```

```
##      Chick          weight           Time       Diet    weight_initial
##  13     : 12   Min.   : 35.0   Min.   : 0.00   1:220   Min.   :39.00
##  9      : 12   1st Qu.: 63.0   1st Qu.: 4.00   2:120   1st Qu.:41.00
##  20     : 12   Median :103.0   Median :10.00   3:120   Median :41.00
##  10     : 12   Mean   :121.8   Mean   :10.72   4:118   Mean   :41.09
##  17     : 12   3rd Qu.:163.8   3rd Qu.:16.00           3rd Qu.:42.00
##  19     : 12   Max.   :373.0   Max.   :21.00           Max.   :43.00
##  (Other):506
##   WeightGroup         Diet4            Diet1            Diet2
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.5294   Mean   :0.2042   Mean   :0.3806   Mean   :0.2076
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##      Diet3
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2076
##  3rd Qu.:0.0000
##  Max.   :1.0000
##
```

```
# get Day 21 and Diet 1+4
Day21 <- subset(chickWeight, Time == 21 & (Diet == 1 | Diet == 4))
summary(Day21$Diet)
```

```
##  1  2  3  4
## 16  0  0  9
```

```
# logit
model <- glm(WeightGroup ~ Diet1, data = Day21, family = "binomial")
summary(model)
```

```
##
## Call:
## glm(formula = WeightGroup ~ Diet1, family = "binomial", data = Day21)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.51678  -0.51678  -0.51678  -0.00008   2.03933
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -19.57    3584.67  -0.005    0.996
## Diet1          17.62    3584.67   0.005    0.996
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13.938  on 24  degrees of freedom
## Residual deviance: 12.057  on 23  degrees of freedom
## AIC: 16.057
##
## Number of Fisher Scoring iterations: 18
```

```
exp(-19.57) #diet4
```

```
## [1] 3.168524e-09
```

```
exp(17.6202-19.57) #diet1
```

```
## [1] 0.1423025
```

```
exp(-19.57)/(1 + exp(-19.57))
```

```
## [1] 3.168524e-09
```

```
exp(17.6202-19.57)/(1 +exp(17.6202-19.57))
```

```
## [1] 0.1245752
```

Answer1(a):

Interpretation:

The model is saying that at Diet = 4, the log odds of a positive outcome(WeightGroup=1) is -19.57. This means the odds of a positive outcome is exp(-19.57) or 3.168524e-09. As a probability, this is 3.168524e-09 / (1 + 3.168524e-09), or about 0. The overall probability of being WeightGroup is 0.

Meanwhile, at Diet = 1, the odds are exp(17.6202-19.57) = 0.1423025, so the probability of a positive outcome is 0.1423025/(1 + 0.1423025) or 0.1245751.

In sum, Diet1 or Diet4 is not significantly associated with WeightGroup on Day21.

Q1(b). Consider comparing Diet Levels 1 and 4 on Day 21.

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

```
# with weight_initial
model2 <- glm(WeightGroup ~ Diet1 +offset(log(weight_initial)), data = Day21, family = "binomia
l")
#Using offset, we are explicitly adjusting fo the influence of initial weight without estimating
a separate coefficient for it. We believe the effect of weight_initial on the log-odds of Weight
Group is known and should not estimated in the model.
#Use offset makes more sense because the question said "adjusting"
summary(model2)
```

```
##
## Call:
## glm(formula = WeightGroup ~ Diet1 + offset(log(weight_initial)),
##     family = "binomial", data = Day21)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.52508  -0.51351  -0.51351  -0.00008   2.04515
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -23.28    3584.33  -0.006    0.995
## Diet1          17.60    3584.33   0.005    0.996
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13.921  on 24  degrees of freedom
## Residual deviance: 12.063  on 23  degrees of freedom
## AIC: 16.063
##
## Number of Fisher Scoring iterations: 18
```

```
model22 <- glm(WeightGroup ~ Diet1 +weight_initial, data = Day21, family = "binomial")
# We estimate coefficient for weight_initial.
#summary(model22) results not significant
```

```
exp(-23.28) #diet4
```

```
## [1] 7.755762e-11
```

```
exp(17.60-23.28) #diet1
```

```
## [1] 0.003413558
```

```
exp(-23.28)/(1+exp(-23.28))
```

```
## [1] 7.755762e-11
```

```
exp(17.60-23.28)/(1+exp(exp(17.60-23.28)))
```

```
## [1] 0.001703866
```

Answer Q1(b):

Interpretation:

By taking offset of weight_initial, the model is saying that at Diet = 4, the log odds of a positive outcome(WeightGroup=1) is -23.28. This means the odds of a positive outcome is exp(-23.28) or 7.755762e-11. As a probability, this is 7.755762e-11 / (1 + 7.755762e-11), or about 0. The overall probability of being WeightGroup is 0.

Meanwhile, at Diet = 1, the odds are exp(17.60-23.28) = 0.003413558, so the probability of a positive outcome is 0.003413558/(1 + 0.003413558) or 0.001703866.

In sum, Diet1 or Diet4 is not significantly associated with WeightGroup on Day21. Even if we are not using offset, but directly adding weight_initial to the model. The result keeps same.

Q2(a). Repeat 1 for all 4 Diet Levels

```
# get Day 21 and Diet 1+4
Day21_only <- subset(chickWeight, Time == 21)
summary(Day21_only$Diet)
```

```
##  1  2  3  4
## 16 10 10  9
```

```
# logit
model3 <- glm(WeightGroup ~ Diet1+Diet2+Diet3, data = Day21_only, family = "binomial")
summary(model3)
```

```
##
## Call:
## glm(formula = WeightGroup ~ Diet1 + Diet2 + Diet3, family = "binomial",
##     data = Day21_only)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.51678  -0.51678  -0.45904  -0.00005   2.14597
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.057e+01  5.910e+03  -0.003    0.997
## Diet1        1.862e+01  5.910e+03   0.003    0.997
## Diet2        1.837e+01  5.910e+03   0.003    0.998
## Diet3       -3.265e-08  8.147e+03   0.000    1.000
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22.044  on 44  degrees of freedom
## Residual deviance: 18.558  on 41  degrees of freedom
## AIC: 26.558
##
## Number of Fisher Scoring iterations: 19
```

```
exp(-2.057e+01) #diet4
```

```
## [1] 1.165635e-09
```

```
exp(1.862e+01-2.057e+01) #diet1
```

```
## [1] 0.1422741
```

```
exp(1.837e+01-2.057e+01) #diet2
```

```
## [1] 0.1108032
```

```
exp(-3.265e-08-2.057e+01) #diet3
```

```
## [1] 1.165635e-09
```

```
0.1422741/(1 + 0.1422741) #diet1
```

```
## [1] 0.1245534
```

```
0.1108032/(1+0.1108032) #diet2
```

```
## [1] 0.09975052
```

Answer Q2(a):

Interpretation:

Odds ratio for Diet1, Diet2, Diet3, Diet4 are 0.1422741, 0.1108032, 1.165635e-09 and 1.165635e-09.

At Diet = 4, the log odds of a positive outcome(WeightGroup=1) is -2.057e+01. This means the odds of a positive outcome is exp(-2.057e+01) or 1.165635e-09. As a probability, this is 1.165635e-09 / (1 + 1.165635e-09), or about 0. The overall probability of being WeightGroup is 0.

At Diet = 1, the odds are exp(1.862e+01-2.057e+01) = 0.1422741, so the probability of a positive outcome is 0.1422741/(1 + 0.1422741) or 0.1245534.

At Diet = 2, the odds are exp(1.837e+01-2.057e+01) = 0.1108032, so the probability of a positive outcome is 0.1108032/(1 + 0.1108032) or 0.09975052.

At Diet = 3, the odds are exp(-3.265e-08-2.057e+01) = 1.165635e-09, so the probability of a positive outcome is 1.165635e-09/(1 + 1.165635e-09) or 0.

Without adjusting for Birth Weight, we get fairly high p-value. There is no significant association between for all 4 Diet levels and WeightGroup on Day21.

Q2(b). Repeat 1 for all 4 Diet Levels adjusting for birth_weight

```
# get Day 21 and Diet 1+4
Day21_only <- subset(chickWeight, Time == 21)
summary(Day21_only)
```

```
##      Chick        weight           Time    Diet   weight_initial
##  13     : 1   Min.   : 74.0   Min.   :21   1:16   Min.   :39.00
##  9      : 1   1st Qu.:167.0   1st Qu.:21   2:10   1st Qu.:41.00
##  20     : 1   Median :205.0   Median :21   3:10   Median :41.00
##  10     : 1   Mean   :218.7   Mean   :21   4: 9   Mean   :41.07
##  17     : 1   3rd Qu.:266.0   3rd Qu.:21          3rd Qu.:42.00
##  19     : 1   Max.   :373.0   Max.   :21          Max.   :43.00
##  (Other):39
##   WeightGroup          Diet4         Diet1            Diet2
##  Min.   :0.00000   Min.   :0.0   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.0   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.00000   Median :0.0   Median :0.0000   Median :0.0000
##  Mean   :0.06667   Mean   :0.2   Mean   :0.3556   Mean   :0.2222
##  3rd Qu.:0.00000   3rd Qu.:0.0   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.00000   Max.   :1.0   Max.   :1.0000   Max.   :1.0000
##
##      Diet3
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.2222
##  3rd Qu.:0.0000
##  Max.   :1.0000
##
```

```
# logit
model32 <- glm(WeightGroup ~ Diet1+Diet2+Diet3+offset(log(weight_initial)), data = Day21_only, f
amily = "binomial")
summary(model32)
```

```
##
## Call:
## glm(formula = WeightGroup ~ Diet1 + Diet2 + Diet3 + offset(log(weight_initial)),
##     family = "binomial", data = Day21_only)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.52508  -0.51351  -0.44986  -0.00005   2.13271
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.428e+01  5.910e+03  -0.004    0.997
## Diet1        1.860e+01  5.910e+03   0.003    0.997
## Diet2        1.837e+01  5.910e+03   0.003    0.998
## Diet3        2.173e-03  8.146e+03   0.000    1.000
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 21.957  on 44  degrees of freedom
## Residual deviance: 18.502  on 41  degrees of freedom
## AIC: 26.502
##
## Number of Fisher Scoring iterations: 19
```

```
model33 <- glm(WeightGroup ~ Diet1+Diet2+Diet3+weight_initial, data = Day21_only, family = "bino
mial")
#summary(model33) results not significant
```

```
exp(-2.428e+01) #diet4
```

```
## [1] 2.853185e-11
```

```
exp(1.862e+01-2.428e+01) #diet1
```

```
## [1] 0.003482517
```

```
exp(1.837e+01-2.428e+01) #diet2
```

```
## [1] 0.002712187
```

```
exp(2.173e-03-2.428e+01) #diet3
```

```
## [1] 2.859392e-11
```

```
0.003482517/(1+0.003482517) #diet1
```

```
## [1] 0.003470431
```

```
0.002712187/(1+0.002712187) #diet2
```

```
## [1] 0.002704851
```

Answer Q2(b):

Interpretation:

Odds ratio for Diet1, Diet2, Diet3, Diet4 are 0.003482517, 0.002712187, 2.859392e-11 and 2.853185e-11.

At Diet = 4, the log odds of a positive outcome(WeightGroup=1) is -2.428e+01. This means the odds of a positive outcome is exp(-2.428e+01) or 2.853185e-11. As a probability, this is 2.853185e-11 / (1 + 2.853185e-11), or about 0. The overall probability of being WeightGroup is 0.

At Diet = 1, the odds are exp(1.862e+01-2.428e+01) = 0.003482517, so the probability of a positive outcome is 0.003482517/(1 + 0.003482517) or 0.003482517

At Diet = 2, the odds are exp(1.837e+01-2.428e+01) = 0.002712187, so the probability of a positive outcome is 0.002712187/(1 + 0.002712187) or 0.002704851

At Diet = 3, the odds are exp(2.173e-03-2.428e+01) = 2.859392e-11, so the probability of a positive outcome is 2.859392e-11(1 + 2.859392e-11) or 0.

With adjusting for Birth Weight, we get fairly high p-value. There is no significant association between for all 4 Diet levels and WeightGroup on Day21. Even if we are not using offset, but directly adding weight_initial to the model. The result keeps same.

Q3(a). Repeat 1 using the L-1 without birth_weight

```
library(Matrix)
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.2.3
```

```
## Loaded glmnet 4.1-8
```

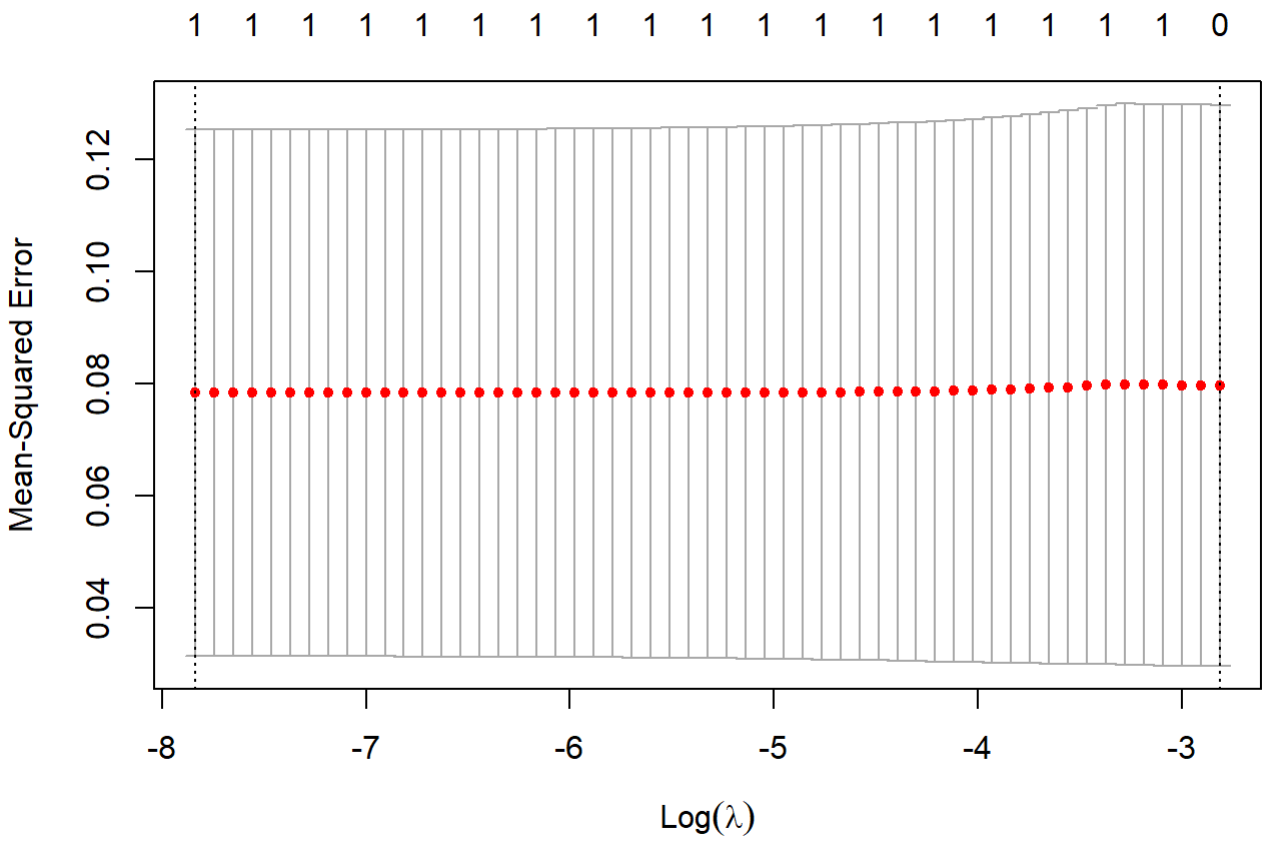
```
Day21 <- subset(chickWeight, Time == 21 & (Diet == 1 | Diet == 4))
X <- model.matrix(WeightGroup ~ Diet - 1, data = Day21)
y <- as.numeric(Day21$WeightGroup)
fit <- glmnet(X, y)
model4 <- cv.glmnet(X, y, alpha = 1)
```

```
## Warning: Option grouped=FALSE enforced in cv.glmnet, since < 3 observations per
## fold
```

```
model4
```

```
##
## Call:  cv.glmnet(x = X, y = y, alpha = 1)
##
## Measure: Mean-Squared Error
##
##      Lambda Index Measure      SE Nonzero
## min 0.00039    55 0.07841 0.04695       1
## 1se 0.06000     1 0.07973 0.05005       0
```

```
plot(model4)
```




```
coef(model4, s = "lambda.min")
```

```
## 5 x 1 sparse Matrix of class "dgCMatrix"
##                       s1
## (Intercept) 0.0005263466
## Diet1       0.1241775835
## Diet2       .
## Diet3       .
## Diet4       .
```

Answer Q3(a)

In our case, “s1” indicates WeightGroup = 1. Baseline Diet4, Diet1 have non-zero coefficients of 0.0005263466 and 0.1241775835.The Diet levels (Diet2, Diet3) have no effect on the outcome as they have zero coefficients.

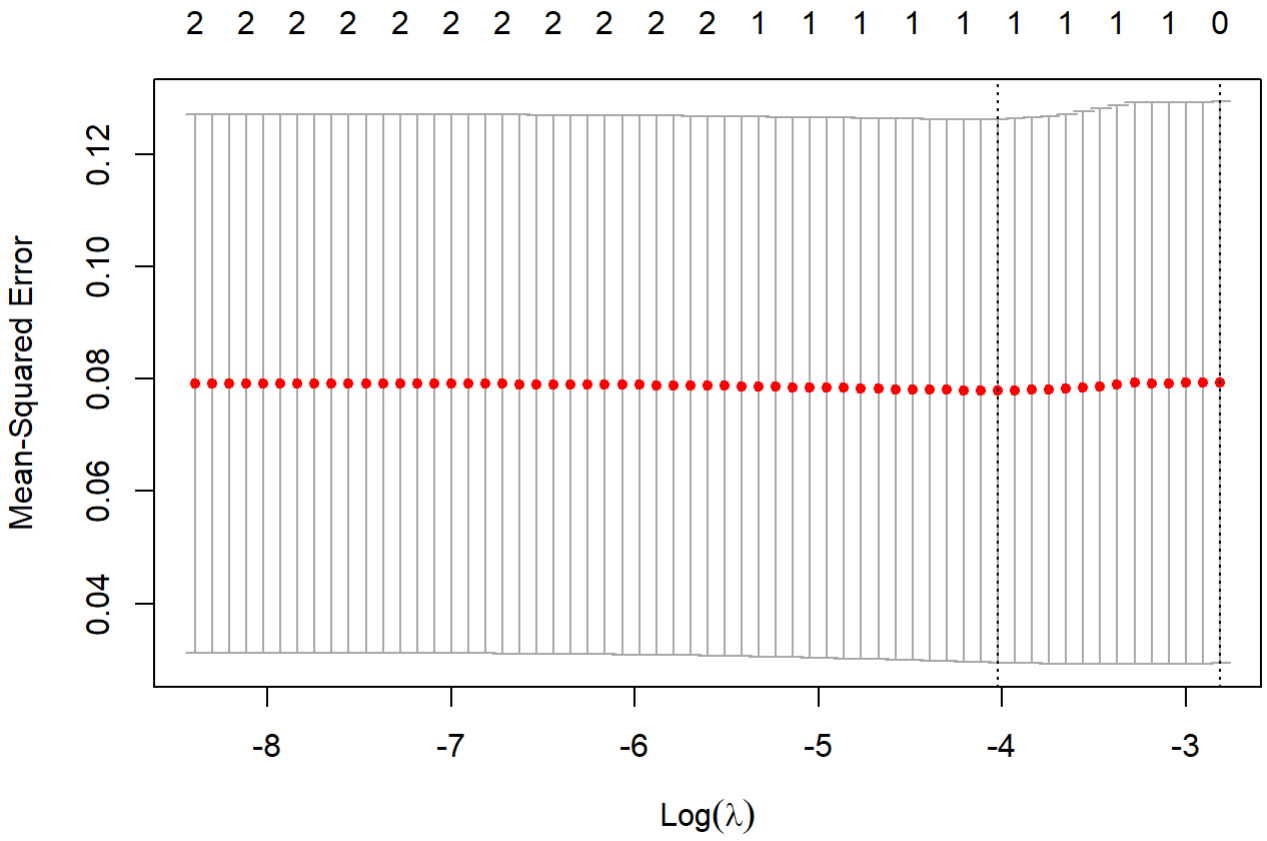
There is two associations. One for Diet4 and WeightGroup, another one for Diet1 and WeightGroup. The Diet levels (Diet2, Diet3) have no effect on the WeightGroup.

Q3(b). Repeat 1 using the L-1 with birth_weight

```
library(Matrix)
library(glmnet)
Day21 <- subset(chickWeight, Time == 21 & (Diet == 1 | Diet == 4))
X <- model.matrix(WeightGroup ~ Diet + weight_initial - 1, data = Day21)
y <- as.numeric(Day21$WeightGroup)
fit <- glmnet(X, y)
model5 <- cv.glmnet(X, y, alpha = 1)
```

```
## Warning: Option grouped=FALSE enforced in cv.glmnet, since < 3 observations per
## fold
```

```
plot(model5)
```




```
coef(model5, s = "lambda.min")
```

```
## 6 x 1 sparse Matrix of class "dgCMatrix"
##                         s1
## (Intercept)    0.02386918
## Diet1          0.08770441
## Diet2          .
## Diet3          .
## Diet4          .
## weight_initial .
```

Answer Q3(b) with adjust

In our case, “s1” indicates WeightGroup = 1. Baseline Diet4 and Diet1 have non-zero coefficients of 0.02619639 and 0.08406814. The Diet levels (Diet2, Diet3) have no effect on the outcome as they have zero coefficients.

There is two associations while adjusting for weight_initial. One for Diet4 and WeightGroup, another one for Diet1 and WeightGroup. The Diet levels (Diet2, Diet3) have no effect on the WeightGroup.