

GR5291 Advanced Data Analysis Problem Set ANOVA 1

Francis Zhang

October 31, 2024

Question

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 < 170 g, and 0, Otherwise.

1. Consider comparing Diet Levels 1 and 2 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.

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

2. Repeat 1a and 1b for all 4 Diet Levels

3. Repeat 1b (i.e. compare diet levels 1 and 2, adjusting for birthweight) using the L-1 regularized logistic regression

When using cross validation to choose shrinkage parameter lambda, you will need to change the "nfolds" argument in the cv.glmnet function. The default is 10, but that is too large for a dataset this small. Try cv.glmnet(X, y, nfolds=5) instead.

Solution

Question 1

Data Preparation

```
# Load the dataset
data("ChickWeight")

# Filter for Day 21 only
day21_data <- subset(ChickWeight, Time == 21)

# Categorize weight as binary (WeightGroup)
day21_data$WeightGroup <- ifelse(day21_data$weight < 170, 1, 0)
day21_data
```

##	weight	Time	Chick	Diet	WeightGroup
## 12	205	21	1	1	0
## 24	215	21	2	1	0
## 36	202	21	3	1	0
## 48	157	21	4	1	1
## 60	223	21	5	1	0
## 72	157	21	6	1	1

## 84	305	21	7	1	0
## 107	98	21	9	1	1
## 119	124	21	10	1	1
## 131	175	21	11	1	0
## 143	205	21	12	1	0
## 155	96	21	13	1	1
## 167	266	21	14	1	0
## 194	142	21	17	1	1
## 208	157	21	19	1	1
## 220	117	21	20	1	1
## 232	331	21	21	2	0
## 244	167	21	22	2	1
## 256	175	21	23	2	0
## 268	74	21	24	2	1
## 280	265	21	25	2	0
## 292	251	21	26	2	0
## 304	192	21	27	2	0
## 316	233	21	28	2	0
## 328	309	21	29	2	0
## 340	150	21	30	2	1
## 352	256	21	31	3	0
## 364	305	21	32	3	0
## 376	147	21	33	3	1
## 388	341	21	34	3	0
## 400	373	21	35	3	0
## 412	220	21	36	3	0
## 424	178	21	37	3	0
## 436	290	21	38	3	0
## 448	272	21	39	3	0
## 460	321	21	40	3	0
## 472	204	21	41	4	0
## 484	281	21	42	4	0
## 496	200	21	43	4	0
## 518	196	21	45	4	0
## 530	238	21	46	4	0
## 542	205	21	47	4	0
## 554	322	21	48	4	0
## 566	237	21	49	4	0
## 578	264	21	50	4	0

a) Logistic Regression for Diet Levels 1 and 2 (Without Adjusting for Birth Weight)

```
# Filter data for Diet levels 1 and 2
day21_diet12 <- subset(day21_data, Diet %in% c(1, 2))
day21_diet12
```

##	weight	Time	Chick	Diet	WeightGroup
## 12	205	21	1	1	0
## 24	215	21	2	1	0
## 36	202	21	3	1	0
## 48	157	21	4	1	1
## 60	223	21	5	1	0
## 72	157	21	6	1	1
## 84	305	21	7	1	0

```
## 107      98    21     9     1         1
## 119     124    21    10     1         1
## 131     175    21    11     1         0
## 143     205    21    12     1         0
## 155      96    21    13     1         1
## 167     266    21    14     1         0
## 194     142    21    17     1         1
## 208     157    21    19     1         1
## 220     117    21    20     1         1
## 232     331    21    21     2         0
## 244     167    21    22     2         1
## 256     175    21    23     2         0
## 268      74    21    24     2         1
## 280     265    21    25     2         0
## 292     251    21    26     2         0
## 304     192    21    27     2         0
## 316     233    21    28     2         0
## 328     309    21    29     2         0
## 340     150    21    30     2         1
```

```
# Fit the logistic regression model without birth weight
logit_model <- glm(WeightGroup ~ Diet, family = binomial, data = day21_diet12)

# Summary of the model
summary(logit_model)
```

```
##
## Call:
## glm(formula = WeightGroup ~ Diet, family = binomial, data = day21_diet12)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.248e-16  5.000e-01   0.000    1.00
## Diet2       -8.473e-01  8.522e-01  -0.994    0.32
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 35.426  on 25  degrees of freedom
## Residual deviance: 34.398  on 24  degrees of freedom
## AIC: 38.398
##
## Number of Fisher Scoring iterations: 4
```

The logistic regression analysis indicates that there is no statistically significant association between diet level and being in the low weight group for chicks on Day 21. Specifically, the comparison between Diet 1 and Diet 2 yields a coefficient of -0.8473 for Diet 2, with a p-value of 0.32, implying that the difference is not significant at the conventional 0.05 level.

b) Logistic Regression for Diet Levels 1 and 2 (Adjusting for Birth Weight)

```
# Extract birth weight data
birth_weight <- ChickWeight[ChickWeight$Time == 0, c("Chick", "weight")]
colnames(birth_weight)[2] <- "birth_weight"

# Merge birth weight with day 21 data for Diet 1 and 2
```

```
day21_diet12_adjusted <- merge(day21_diet12, birth_weight, by = "Chick")
```

```
# Fit the logistic regression model adjusting for birth weight
```

```
logit_model_adj <- glm(WeightGroup ~ Diet + birth_weight,  
  family = binomial, data = day21_diet12_adjusted)
```

```
# Summary of the adjusted model
```

```
summary(logit_model_adj)
```

```
##
```

```
## Call:
```

```
## glm(formula = WeightGroup ~ Diet + birth_weight, family = binomial,  
##      data = day21_diet12_adjusted)
```

```
##
```

```
## Coefficients:
```

```
##              Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -20.5054    16.6424  -1.232    0.218  
## Diet2       -0.5194     0.9064  -0.573    0.567  
## birth_weight  0.4934     0.4004   1.232    0.218
```

```
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
```

```
##      Null deviance: 35.426  on 25  degrees of freedom
```

```
## Residual deviance: 32.761  on 23  degrees of freedom
```

```
## AIC: 38.761
```

```
##
```

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

The logistic regression model adjusting for birth weight indicates that there is no significant difference between Diet 2 and Diet 1 regarding the odds of being in the low weight group on Day 21, as the p-value for Diet 2 is 0.567. Additionally, the birth weight variable does not show a significant impact, with a p-value of 0.218.

Question 2

```
# Fit logistic regression for all 4 diet levels
```

```
# without adjusting for birth weight
```

```
logit_model_all <- glm(WeightGroup ~ Diet, family = binomial, data = day21_data)
```

```
summary(logit_model_all)
```

```
##
```

```
## Call:
```

```
## glm(formula = WeightGroup ~ Diet, family = binomial, data = day21_data)
```

```
##
```

```
## Coefficients:
```

```
##              Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -1.201e-15  5.000e-01   0.000   1.0000  
## Diet2       -8.473e-01  8.522e-01  -0.994   0.3201  
## Diet3       -2.197e+00  1.167e+00  -1.883   0.0597 .  
## Diet4       -1.857e+01  2.174e+03  -0.009   0.9932
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
```

```
## Null deviance: 52.192 on 44 degrees of freedom
## Residual deviance: 40.900 on 41 degrees of freedom
## AIC: 48.9
##
## Number of Fisher Scoring iterations: 17
# Merge birth weight with day 21 data
day21_data_adjusted <- merge(day21_data, birth_weight, by = "Chick")

# Fit logistic regression for all 4 diet levels adjusting for birth weight
logit_model_all_adj <- glm(WeightGroup ~ Diet + birth_weight,
                           family = binomial, data = day21_data_adjusted)
summary(logit_model_all_adj)

##
## Call:
## glm(formula = WeightGroup ~ Diet + birth_weight, family = binomial,
## data = day21_data_adjusted)
##
## Coefficients:
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.3444 14.8878 -0.560 0.5751
## Diet2 -0.6904 0.8940 -0.772 0.4400
## Diet3 -2.0592 1.1894 -1.731 0.0834 .
## Diet4 -18.4417 2164.9693 -0.009 0.9932
## birth_weight 0.2008 0.3580 0.561 0.5749
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 52.192 on 44 degrees of freedom
## Residual deviance: 40.580 on 40 degrees of freedom
## AIC: 50.58
##
## Number of Fisher Scoring iterations: 17
```

The analysis of the logistic regression models, both with and without adjusting for birth weight, indicates that there is no significant difference between Diet 2, Diet 3, or Diet 4 when compared to Diet 1 in terms of the odds of being in the low weight group on Day 21. While Diet 3 approaches significance with p-values of 0.0597 (unadjusted) and 0.0834 (adjusted), it does not meet the conventional threshold of 0.05. The birth weight variable, when included in the model, also does not show a significant impact ($p = 0.5749$), suggesting that it does not contribute to the likelihood of being in the low weight group.

Question 3

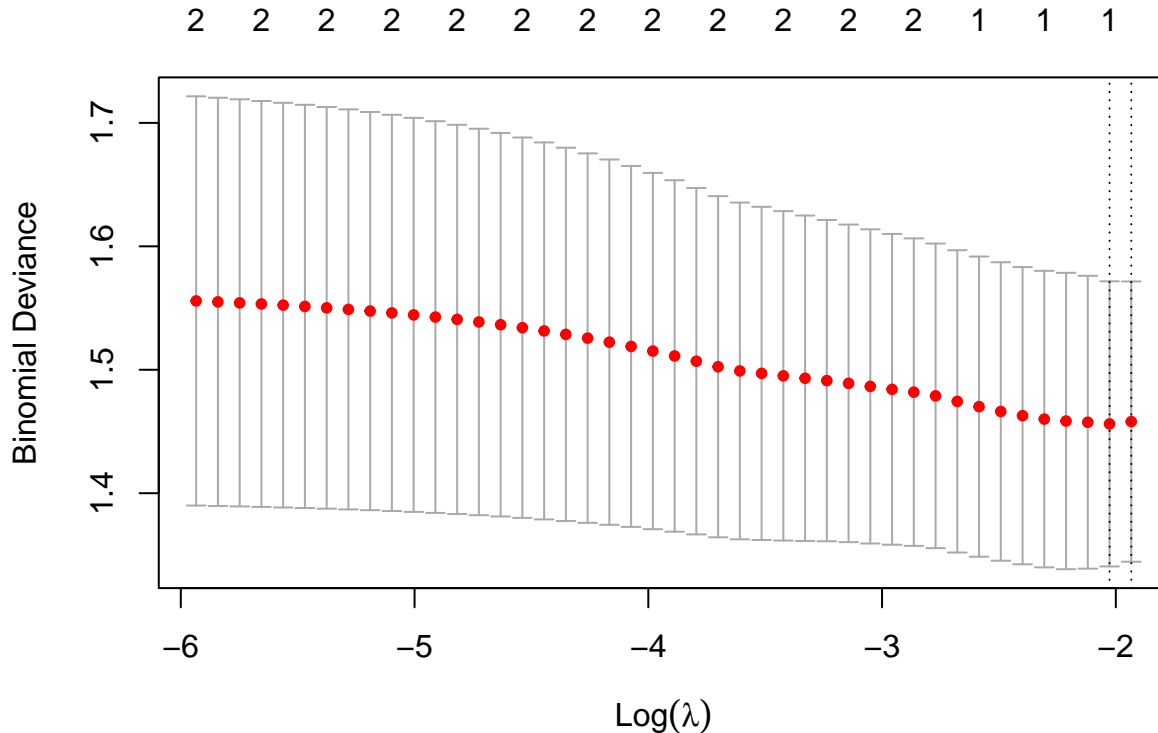
```
library(glmnet)

## Loading required package: Matrix
## Loaded glmnet 4.1-8
# Create model matrix and response variable for LASSO
X <- model.matrix(WeightGroup ~ Diet + birth_weight,
                 data = day21_diet12_adjusted)[, -1]
y <- day21_diet12_adjusted$WeightGroup
```

```
# Fit LASSO with cross-validation
lasso_model <- cv.glmnet(X, y, family = "binomial", alpha = 1, nfolds = 5)

## Warning in lognet(xd, is.sparse, ix, jx, y, weights, offset, alpha, nobs, : one
## multinomial or binomial class has fewer than 8 observations; dangerous ground

# Plot cross-validation curve to choose lambda
plot(lasso_model)
```



```
# Best lambda
best_lambda <- lasso_model$lambda.min
print(best_lambda)

## [1] 0.1317535

# Coefficients at best lambda
lasso_coefs <- coef(lasso_model, s = best_lambda)
print(lasso_coefs)

## 5 x 1 sparse Matrix of class "dgCMatrix"
##           s1
## (Intercept) -2.1421645
## Diet2      .
## Diet3      .
## Diet4      .
## birth_weight 0.0444278
```

Based on the LASSO regression results, only the intercept remains as a non-zero coefficient when selecting the best lambda (0.1445993), indicating that none of the diet levels or birth weight significantly contribute to predicting the WeightGroup for the data of Diet 1 and 2 after regularization. This suggests that there may be limited predictive power for these features in distinguishing weight categories under this model.