

STAT6420  
Final Project Cover Page    December 5, 2023

**Name : Agbolade Akande**

Pledge:

I have neither given nor received any unauthorized aid for this project. I abide by the academic honor code of UGA.

## Contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>Exploratory Data Analyses</b>	<b>4</b>
2.1	First research question . . . . .	4
2.2	Second research question . . . . .	6
2.3	Third research question . . . . .	8
2.4	Fifth research question . . . . .	10
<b>3</b>	<b>Statistical Analyses</b>	<b>11</b>
3.1	First research question . . . . .	11
3.2	Second research question . . . . .	11
3.3	Fourth research question . . . . .	12
3.4	Fifth research question . . . . .	12
<b>4</b>	<b>Conclusions</b>	<b>13</b>
<b>5</b>	<b>Appendix: R code</b>	<b>14</b>

## List of Figures

1	Residual vs fitted plot of regular linear regression . . . . .	7
2	Deviance plot of logistic regression . . . . .	8
3	Residual vs Fitted Plot and Boxcox Plot of Linear regression . . . . .	9
4	Fitted response vs original response for both linear regression fit and gamma regression fit . . . . .	9
5	Leverage Plot and Cooks Distance Plot. The red dots are the entries 137, 385 and 1541. . . . .	10
6	Cooks distances when the exceptional entries are removed . . . . .	10

## List of Tables

1	Confusion matrix for logistic regression . . . . .	13
2	Confusion matrix for logistic regression with cross-validation . . . . .	13

## Summary

In this report, we will do some analysis on the dataset comprising of samples taken in different farms from different animal sources. We would like to answer these five questions using generalized linear regression analysis:

1. Are `Camplyobacter`, `Salmonella`, and `Listeria` predictive of each other.
2. Model the probability of `Camplyobacter` in terms of the `Type` and `Month` variable
3. Model the variable `EcoliLog10` in terms of the `Type` and `Month` variable
4. Are `Farm` and `Flock` needed to be controlled to answer the previous two research questions.
5. Can both the `SampleType` and `PastureTime` be predicted using the proportion of bacteria found? If not, can the `SampleType` alone be predicted using using the proportion of bacteria found?

After analyzing the data, setting up some models and perform some specific test, we arrived at these conclusions for our questions: With our results, we can finally answer our questions presented in the Introduction.

1. We do not have enough information to show that `Camplyobacter`, `Salmonella`, and `Listeria` are predictive of each other.
2. A model was fitted and we get a dispersion effect of approximately 0.89. So there is lack of overdispersion but there is sign of underdispersion.
3. A model was fitted and it was shown that there are no influential points in our dataset. So those problematic points found can be considered to be outliers.
4. The `Farm` and `Flock` should not be controlled to answer the previous questions because the predictors used in the previous questions are not consistent in every `Farm` and every `Flock`.
5. We could not predict both the `SampleType` and `PastureTime` variables in unison. Rather, we predicted only the `SampleType` using the proportion of the bacteria samples found. Using the entire dataset, we have a small classification error rate of 6.45%. Using cross-validation, we get an average classification error rate of 7%.

# 1 Introduction

In this project, we will study the dataset that contains information on the. The goal of these report is to answer these five questions

1. Are Camplyobacter, Salmonella, and Listeria predictive of each other.
2. Model the probability of Camplyobacter in terms of the Type and Month variable
3. Model the variable EcoliLog10 in terms of the Type and Month variable
4. Are Farm and Flock needed to be controlled to answer the previous two research questions.
5. Can both the SampleType and PastureTime be predicted using the proportion of bacteria found?  
If not, can the SampleType alone be predicted using using the proportion of bacteria found?

## 2 Exploratory Data Analyses

### 2.1 First research question

We want to see the predictive power of Camplyobacter, Salmonella, and Listeria with each other. Hence, we will make three models as follows

$$\text{Camplyobacter} = \beta_0 + \beta_1 \cdot \text{Salmonella} + \beta_2 \cdot \text{Listeria}$$

$$\text{Salmonella} = \beta_0 + \beta_1 \cdot \text{Camplyobacter} + \beta_2 \cdot \text{Listeria}$$

$$\text{Listeria} = \beta_0 + \beta_1 \cdot \text{Salmonella} + \beta_2 \cdot \text{Camplyobacter}$$

The summary of these models are shown below:

```
> summary(model_Camplyobact)

Call:
lm(formula = "Camplyobact~Listeria+Salmonella", data = broilers)

Residuals:
    Min       1Q   Median       3Q      Max
-0.4676 -0.4405 -0.4395  0.5595  0.5605

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4404735   0.0134922  32.647  <2e-16 ***
Listeria     -0.0009209   0.0320169  -0.029   0.977
Salmonella    0.0271414   0.0303931   0.893   0.372
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4973 on 1887 degrees of freedom
(2 observations deleted due to missingness)
Multiple R-squared:  0.0004237, Adjusted R-squared:  -0.0006358
```

F-statistic: 0.3999 on 2 and 1887 DF, p-value: 0.6704

```
> summary(model_Salmonella)
```

Call:

```
lm(formula = "Salmonella~Campylobact+Listeria", data = broilers)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.1830	-0.1830	-0.1674	-0.1601	0.8555

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.16742	0.01219	13.739	<2e-16 ***
Campylobact	0.01556	0.01743	0.893	0.372
Listeria	-0.02291	0.02424	-0.945	0.345

---

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

Residual standard error: 0.3766 on 1887 degrees of freedom

(2 observations deleted due to missingness)

Multiple R-squared: 0.0008964, Adjusted R-squared: -0.0001626

F-statistic: 0.8465 on 2 and 1887 DF, p-value: 0.4291

```
> summary(model_Listeria)
```

Call:

```
lm(formula = "Listeria~Salmonella+Campylobact", data = broilers)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.1540	-0.1540	-0.1535	-0.1333	0.8671

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.1540067	0.0116058	13.270	<2e-16 ***
Salmonella	-0.0206572	0.0218524	-0.945	0.345
Campylobact	-0.0004761	0.0165519	-0.029	0.977

---

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

Residual standard error: 0.3575 on 1887 degrees of freedom

(2 observations deleted due to missingness)

Multiple R-squared: 0.0004746, Adjusted R-squared: -0.0005848

F-statistic: 0.448 on 2 and 1887 DF, p-value: 0.639

We see that the predictors in all the models are not significant which shows signs of colinearity so we will check the correlation matrix.

```
> cor(broilers[,c("Campylobact", "Listeria", "Salmonella")],  
+      use = "complete.obs")  
          Campylobact    Listeria    Salmonella
```

```

Campylobact  1.0000000000 -0.001109826  0.02057247
Listeria     -0.001109826  1.0000000000 -0.02177444
Salmonella   0.020572470 -0.021774441  1.000000000

```

We see that the correlations between different bacteria samples are very low so signs are slim.

To also consider all possible cases, we will consider a logistic model for each of the

## 2.2 Second research question

We will study the effect on the sample type and the months on the presence/absence of *Campylobacter*. We first fit a linear model without interaction terms because there will be too many characters ( $4 \times 7 = 28$ ). The summary of the model is shown below:

```

Call:
lm(formula = "Campylobact~factor(SampleType)+factor(Month) ",
    data = broilers)

Residuals:
    Min       1Q   Median       3Q      Max
-1.01568 -0.21013 -0.01568  0.26482  0.99110

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.13945    0.04212  27.052 < 2e-16 ***
factor(SampleType)Feces -0.23628    0.03053  -7.738 1.64e-14 ***
factor(SampleType)Soil  -0.70929    0.03053 -23.229 < 2e-16 ***
factor(SampleType)WCR-F -0.91052    0.03704 -24.582 < 2e-16 ***
factor(SampleType)WCR-P -0.85714    0.03695 -23.199 < 2e-16 ***
factor(Month)5        -0.12377    0.03950  -3.133  0.00176 **
factor(Month)6        -0.19727    0.03705  -5.324 1.13e-07 ***
factor(Month)7        -0.22003    0.03754  -5.862 5.40e-09 ***
factor(Month)8        -0.23105    0.03917  -5.898 4.34e-09 ***
factor(Month)9        -0.24527    0.04029  -6.088 1.38e-09 ***
factor(Month)10       -0.16800    0.04007  -4.193 2.88e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3786 on 1879 degrees of freedom
(2 observations deleted due to missingness)
Multiple R-squared:  0.423, Adjusted R-squared:  0.4199
F-statistic: 137.8 on 10 and 1879 DF, p-value: < 2.2e-16

```

We see that all of the predictors are significant but the  $R^2$  is moderately low. We check the residual vs fitted plot in Figure 1.

We see in Figure 1 that there is a lack of a horizontal band but rather a downward pattern shown. So we will also consider logistic regression (not probit because we are considering odds). The summary is shown below:

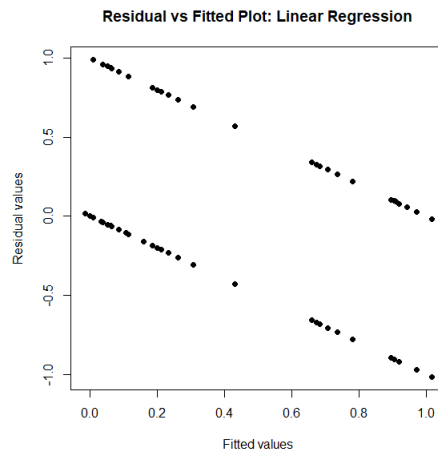


Figure 1: Residual vs fitted plot of regular linear regression

```
Call:
glm(formula = "Campylobact~factor(SampleType)+factor(Month)",
     family = binomial(link = "logit"), data = broilers)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)      3.8825     0.3640  10.667 < 2e-16 ***
factor(SampleType)Feces -1.9172     0.3123  -6.140 8.27e-10 ***
factor(SampleType)Soil -4.0581     0.3151 -12.881 < 2e-16 ***
factor(SampleType)WCR-F -6.5206     0.5428 -12.012 < 2e-16 ***
factor(SampleType)WCR-P -5.2329     0.3883 -13.477 < 2e-16 ***
factor(Month)5      -0.5767     0.2534  -2.276 0.02287 *
factor(Month)6      -1.0494     0.2396  -4.380 1.19e-05 ***
factor(Month)7      -1.2041     0.2430  -4.954 7.26e-07 ***
factor(Month)8      -1.3088     0.2603  -5.027 4.97e-07 ***
factor(Month)9      -1.3930     0.2668  -5.221 1.78e-07 ***
factor(Month)10     -0.8052     0.2676  -3.009 0.00262 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2597.2  on 1889  degrees of freedom
Residual deviance: 1670.0  on 1879  degrees of freedom
(2 observations deleted due to missingness)
AIC: 1692

Number of Fisher Scoring iterations: 6

We will check the deviance plot to see if a transformation is required.
```

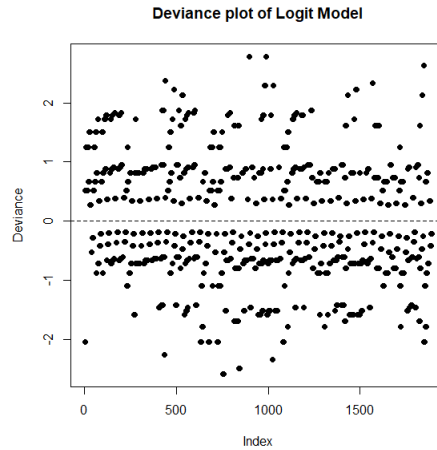


Figure 2: Deviance plot of logistic regression

We see in Figure 2 that the points are distributed evenly so there is no need for a transformation.

### 2.3 Third research question

We will study the effect on the sample type and the animal source on the logarithmic scale of the *E. coli* concentration in the sample. We first fit a linear model with interaction terms and the `EcoliLog10` is shifted by 0.1 so the boxcox plot can be shown. The summary is shown below:

Call:

```
lm(formula = "I(EcoliLog10 + 0.1) ~ factor(SampleType)+ factor(AnimalSource)
+ factor(SampleType)*factor(AnimalSource)", data = fecalsoil)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-5.7246	-0.8384	0.0370	0.9811	4.3188

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	6.63741	0.05691	116.629	< 2e-16 ***
factor(SampleType) Soil	-2.11200	0.08061	-26.200	< 2e-16 ***
factor(AnimalSource) Cattle	-1.68056	0.22025	-7.630	3.98e-14 ***
factor(AnimalSource) Layer	0.59819	0.16086	3.719	0.000207 ***
factor(AnimalSource) Swine	-0.81281	0.20972	-3.876	0.000111 ***
factor(SampleType) Soil:factor(AnimalSource) Cattle	-0.90452	0.31151	-2.904	0.003739 **
factor(SampleType) Soil:factor(AnimalSource) Layer	-1.26201	0.22753		
factor(SampleType) Soil:factor(AnimalSource) Swine	-2.41290	0.29662		



```

factor(SampleType) Soil:factor(AnimalSource) Layer    -5.547  3.40e-08 ***
factor(SampleType) Soil:factor(AnimalSource) Swine   -8.135  8.16e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 1.427 on 1616 degrees of freedom  
 (8 observations deleted due to missingness)  
 Multiple R-squared: 0.5042, Adjusted R-squared: 0.5021  
 F-statistic: 234.8 on 7 and 1616 DF, p-value: < 2.2e-16

We see that the interaction terms are very significant. We see that there is not much fanning in or out in the residual plot and the boxcox plot shows that there is no need for a transformation in Figure 3.

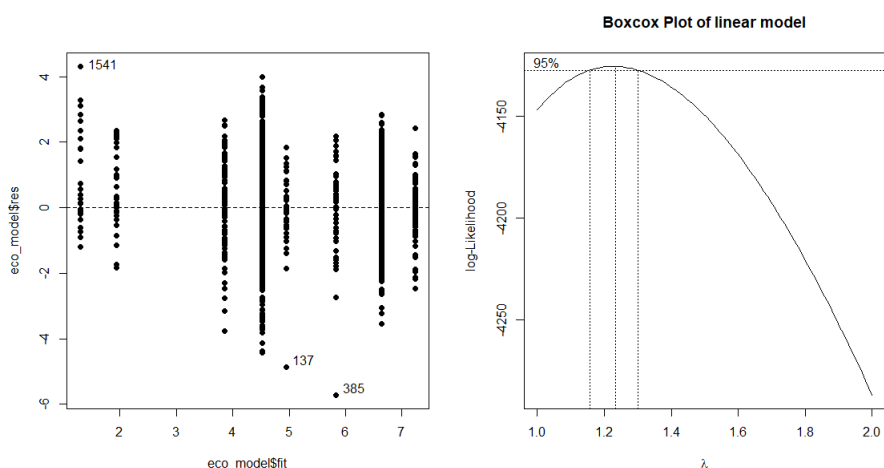


Figure 3: Residual vs Fitted Plot and Boxcox Plot of Linear regression

A Gamma model can be considered but we see from Figure 4, a fitted Gamma model and the linear model have the same fit.

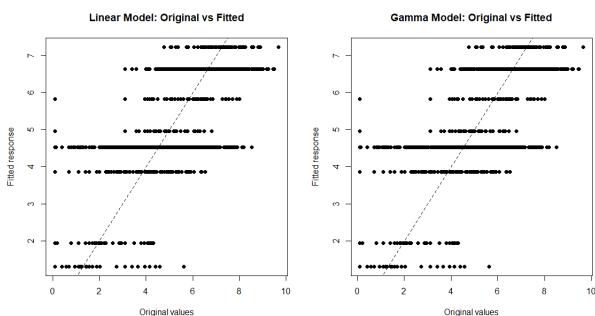


Figure 4: Fitted response vs original response for both linear regression fit and gamma regression fit

So we will stick with the linear model we fitted. Hence, we will check the leverage plot and the cooks distance plot: We see in Figures 3 and 5 that entry 385, 137 and 1541 have high residuals, high leverages

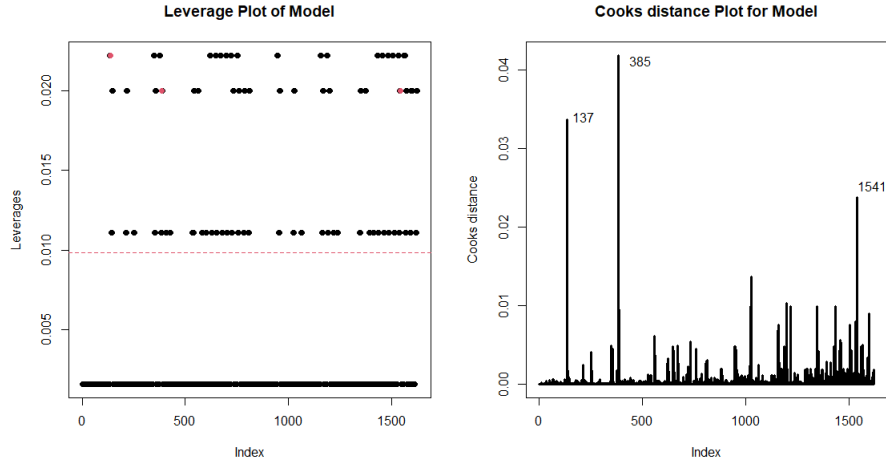


Figure 5: Leverage Plot and Cooks Distance Plot. The red dots are the entries 137, 385 and 1541.

and high cooks distances relative to the other entries. So they are considered to be either outliers or influential points. For more information, we will check the cooks distance if we remove each entry from the model.

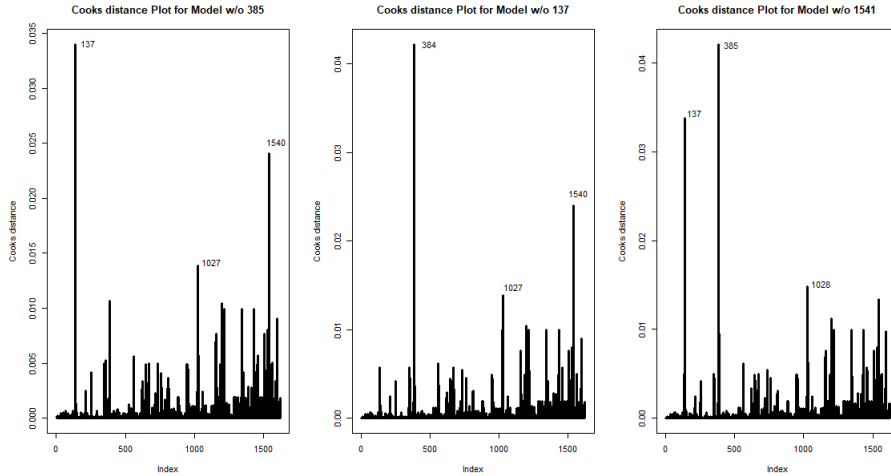


Figure 6: Cooks distances when the exceptional entries are removed

We see that not much changes in the cooks distance, so those entries are not considered to be influential points. So there are no strong influential points in this dataset.

## 2.4 Fifth research question

The question is if we can predict both the Sample Type and Pasture time by using the proportions of bacteria found (A, B, C, D, E). We cannot predict both the Sample Type and the Pasture time because they are two different groups which could be independent. However, we can predict the Sample Type because there are only two variable in the Sample Type so we can use logistic regression. We only need four predictors

because  $A + B + C + D + E = 1$ . If we run all four possible combinations of the four parameters, we see that the deviance for all of them are significant so we will choose the model that has the most significant predictors. The summary of our chosen model is shown below:

```
Call:
glm(formula = Sample_dummy ~ A + B + C + D, family = binomial)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    50.70      12.00   4.225 2.39e-05 ***
A             -55.60      12.20  -4.557 5.19e-06 ***
B             -50.93      12.28  -4.149 3.34e-05 ***
C             -43.03      12.64  -3.404 0.000665 ***
D             -53.97      19.27  -2.801 0.005098 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 300.711  on 216  degrees of freedom
Residual deviance:  69.343  on 212  degrees of freedom
AIC: 79.343

Number of Fisher Scoring iterations: 8
```

### 3 Statistical Analyses

#### 3.1 First research question

Let  $\beta_{Campylobacter}$ ,  $\beta_{Salmonella}$  and  $\beta_{Listeria}$  be the coefficients of the respoective predictors. Our hypotheses are as follow:

$$H_0 : \beta_{Campylobacter} = \beta_{Salmonella} = \beta_{Listeria} = 0$$

$$H_A : \text{at least one of these parameters is } 0$$

We will use is a  $t$ -test in our process using the models we prepared. As we see in both the linear and logistic models, none of the predictors are significant so we fail to reject the null hypothesis. Hence, we do not have enough to show that *Camplyobacter*, *Salmonella*, and *Listeria* are predictive of each other.

#### 3.2 Second research question

We will check for overdispersion in the binomal model fitted. The calculation of the dispersion factor is as follows:

$$\frac{\text{Deviance}}{\text{Error degree of freedom}}$$

Since this is a binomial model, the dispersion factor is expected to be very close to 1. Hence, the dispersion factor of our model is shown below:

```
> campyl_model_binom_logit$deviance/campyl_model_binom_logit$df.residual
[1] 0.8887724
```

The dispersion factor is smaller than 1 so there is no sign of overdispersion but a small sign of underdispersion.

### 3.3 Fourth research question

The models used in the second and third response should not be controlled for the `Farm` and `Flock` variables because it is possible that the predictors used in the previous research questions might not be in each `Farm` or `Flock`. For instance, we will predict the presence of `Campylobacter` using the `Type` and `Month` but only for the `Farm` labeled as `A`.

```
Call:
glm(formula = "Campylobact~factor(SampleType)+factor(Month)",
     family = binomial(link = "logit"), data = broilers[broilers$Farm ==
"A", ])
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )	
(Intercept)	6.5748	1.1852	5.547	2.90e-08	***
factor(SampleType)Feces	-1.6478	1.0649	-1.547	0.121765	
factor(SampleType)Soil	-5.1429	1.0502	-4.897	9.73e-07	***
factor(SampleType)WCR-F	-22.9223	883.4291	-0.026	0.979300	
factor(SampleType)WCR-P	-8.2900	1.4567	-5.691	1.26e-08	***
factor(Month)5	-0.8190	0.7092	-1.155	0.248167	
factor(Month)6	-2.3167	0.6347	-3.650	0.000262	***
factor(Month)7	-2.0530	0.6532	-3.143	0.001672	**
factor(Month)8	-3.2910	0.7047	-4.670	3.01e-06	***
factor(Month)9	-3.1195	0.7630	-4.088	4.34e-05	***

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 619.91 on 449 degrees of freedom  
Residual deviance: 256.00 on 440 degrees of freedom  
(2 observations deleted due to missingness)  
AIC: 276

Number of Fisher Scoring iterations: 17

We see Farm A lacks the 10 Month data so there will be a vast difference in the estimate of the other parameters.

### 3.4 Fifth research question

We will check the confusion matrix of our results. The matrix is shown in table 1:

Hence, the error rate can be evaluated as  $\frac{10+4}{107+10+4+96} = \frac{14}{217} = 6.45\%$ .

Now we will perform cross-validation on our dataset: breaking our training set and test set by 70 to 30. The confusion matrix for our cross-validation is shown in Table 2

We see that our error rate is  $\frac{1+1}{40+1+1+30} = \frac{2}{72} \approx 2.78\%$ .

		Actual values	
		Fecal	Soil
Predicted values	Fecal	107	10
	Soil	4	96

Table 1: Confusion matrix for logistic regression

		Actual values	
		Fecal	Soil
Predicted values	Fecal	40	1
	Soil	1	30

Table 2: Confusion matrix for logistic regression with cross-validation

If we repeat the cross-validation 10 times, we get this list of error rates, the summary of our list is shown below:

```
> summary(1-accuracy_score_list)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.04167 0.05903 0.07639 0.07083 0.08333 0.09722
```

We see that the error rate average and median is around 7% and the error rate does not get larger than 10%.

## 4 Conclusions

With our results, we can finally answer our questions presented in the Introduction.

1. We do not have enough information to show that `Campylobacter`, `Salmonella`, and `Listeria` are predictive of each other.
2. A model was fitted and we get a dispersion effect of approximately 0.89. So there is lack of overdispersion but there is sign of underdispersion.
3. A model was fitted and it was shown that there are no influential points in our dataset. So those problematic points found can be considered to be outliers.
4. The Farm and Flock should not be controlled to answer the previous questions because the predictors used in the previous questions are not consistent in every Farm and every Flock.
5. We could not predict both the `SampleType` and `PastureTime` variables in unison. Rather, we predicted only the `SampleType` using the proportion of the bacteria samples found. Using the entire dataset, we have a small classification error rate of 6.45%. Using cross-validation, we get an average classification error rate of 7%.

## 5 Appendix: R code

```
1 library(readxl)
2 library(faraway)
3 library(MASS)
4 library(leaps)
5
6 dev.new()
7
8 broilers <- read_excel("PasturedPoultryFarms.xlsx", sheet = "Broilers")
9
10 fecalsoil <- read_excel("PasturedPoultryFarms.xlsx", sheet =
  ↳ "FecalSoil")
11
12 Compositional <- read_excel("PasturedPoultryFarms.xlsx", sheet =
  ↳ "Compositional")
13
14
15 #####
16 #
17   ↳ #
18 #
19   ↳ #
20 #
21   ↳ #
22 #####
23
24 model_Campylobact <- lm("Campylobact~Listeria+Salmonella", data =
  ↳ broilers)
25 model_Salmonella <- lm("Salmonella~Campylobact+Listeria", data =
  ↳ broilers)
26 model_Listeria <- lm("Listeria~Salmonella+Campylobact", data = broilers)
27
28 summary(model_Campylobact)
29 summary(model_Salmonella)
30 summary(model_Listeria)
31
32 model_Campylobact_binom <- glm("Campylobact~Listeria+Salmonella", data =
  ↳ broilers,
33                               family = binomial)
34 model_Salmonella_binom <- glm("Salmonella~Campylobact+Listeria", data =
  ↳ broilers,
```

```

34         family = binomial)
35 model_Listeria_binom <- glm("Listeria~Salmonella+Campylobact",data =
  ↳ broilers,
36         family = binomial)
37
38 cor(broilers[,c("Campylobact", "Listeria", "Salmonella")],
39     use = "complete.obs")
40
41
42 summary(model_Campylobact_binom)
43 summary(model_Salmonella_binom)
44 summary(model_Listeria_binom)
45
46
47
48
49 #####
50 #
  ↳ #
51 #
  ↳ #
52 #                               Second Ques
  ↳ #
53 #
  ↳ #
54 #
  ↳ #
55 #####
56
57
58 campyl_model <- lm("Campylobact~factor(SampleType)+factor(Month)",
59                   data = broilers)
60 summary(campyl_model)
61 plot(campyl_model$fit, campyl_model$res,pch = 19,xlab = "Fitted
  ↳ values",
62      ylab = "Residual values",main = "Residual vs Fitted Plot: Linear
  ↳ Regression")
63
64 boxcox(lm("I (Campylobact+0.1)~factor(SampleType)+factor(Month)",
65          data = broilers))
66
67 campyl_model_1 <-
  ↳ lm("log (Campylobact+0.1)~factor(SampleType)+factor(Month)",
68      data = broilers)
69 summary(campyl_model_1)

```

```

70
71 plot(campyl_model_1$fit, campyl_model_1$res, pch = 19)
72
73
74 campyl_model_binom_logit <-
  ↪ glm("Campylobact~factor(SampleType)+factor(Month)",
75       data = broilers, family = binomial(link =
  ↪ "logit"))
76 summary(campyl_model_binom_logit)
77
78
79 plot(residuals(campyl_model_binom_logit, type = "deviance"), pch = 19,
80      ylab="Deviance", main = "Deviance plot of Logit Model")
81 abline(h = 0, lty = 2)
82
83 #Dispersion factor
84 campyl_model_binom_logit$deviance/campyl_model_binom_logit$df.residual
85
86 #####
87 #
88 ↪ #
89 #
90 ↪ #
91 #
92 ↪ #
93 #####
94
95
96 fecalsoil$EcoliLog10 <- as.numeric(fecalsoil$EcoliLog10)
97
98 eco_model <- lm("I(EcoliLog10 + 0.1) ~ factor(SampleType)+
  ↪ factor(AnimalSource) + factor(SampleType)*factor(AnimalSource)",
99      data = fecalsoil)
100 summary(eco_model)
101
102
103 par(mfrow = c(1,2))
104 plot(eco_model$fit, eco_model$res, pch = 19)
105 abline(h=0, lty = 2)
106 identify(eco_model$fit, eco_model$res, atpen = T, tolerance = 0.5)

```



```

107
108 boxcox(eco_model, lambda = seq(1,2,by = 0.05))
109 title("Boxcox Plot of linear model")
110 par(mfrow = c(1,1))
111
112 step(eco_model)
113
114 Cpplot(leaps(model.matrix(eco_model)[,-1],na.omit(fecalsoil$EcoliLog10)+0.1))
115
116 maxadjr(leaps(model.matrix(eco_model)[,-1],na.omit(fecalsoil$EcoliLog10)+0.1,
117             method="adjr2"), best = 8)
118
119 eco_model_gamma <- glm("I(EcoliLog10+0.1) ~ factor(SampleType)+
120   ↪ factor(AnimalSource) + factor(SampleType)*factor(AnimalSource)",
121   data = fecalsoil, family = Gamma(link = "inverse"))
122
123 summary(eco_model_gamma)
124
125 par(mfrow = c(1,2))
126 plot(na.omit(fecalsoil$EcoliLog10)+0.1,eco_model$fit,pch = 19,
127       xlab = "Original values",ylab="Fitted response",main = "Linear
128   ↪ Model: Original vs Fitted")
129
130 abline(a=0,b=1,lty = 2)
131
132 plot(na.omit(fecalsoil$EcoliLog10)+0.1,eco_model_gamma$fit,pch = 19,
133       xlab = "Original values",ylab="Fitted response",main = "Gamma
134   ↪ Model: Original vs Fitted")
135
136 abline(a=0,b=1,lty = 2)
137 par(mfrow = c(1,1))
138
139 par(mfrow = c(1,2))
140 plot(hat(model.matrix(eco_model)),pch = 19,ylab = 'Leverages',
141       main = 'Leverage Plot of Model')
142
143 abline(h =
144   ↪ 2*ncol(model.matrix(eco_model))/nrow(model.matrix(eco_model)),lty =
145   ↪ 2, col = 2)
146
147 points(137,hat(model.matrix(eco_model))[137], col = 2,pch = 19)
148 points(385,hat(model.matrix(eco_model))[385], col = 2,pch = 19)
149 points(1541,hat(model.matrix(eco_model))[1541], col = 2,pch = 19)
150
151
152 plot(cooks.distance(eco_model),type = 'h',lwd = 3,ylab = 'Cooks
153   ↪ distance',
154       main = 'Cooks distance Plot for Model')

```

```

146 identify(1:nrow(model.matrix(eco_model)), cooks.distance(eco_model),
147           tolerance = 0.5, atpen = TRUE)
148 par(mfrow = c(1,1))
149
150 eco_model_385 <- lm("I(EcoliLog10 + 0.1) ~ factor(SampleType)+
151   ↪ factor(AnimalSource) + factor(SampleType)*factor(AnimalSource)",
152                       data = na.omit(fecalsoil)[-385,])
153
154 eco_model_137 <- lm("I(EcoliLog10 + 0.1) ~ factor(SampleType)+
155   ↪ factor(AnimalSource) + factor(SampleType)*factor(AnimalSource)",
156                       data = na.omit(fecalsoil)[-137,])
157
158 eco_model_1541 <- lm("I(EcoliLog10 + 0.1) ~ factor(SampleType)+
159   ↪ factor(AnimalSource) + factor(SampleType)*factor(AnimalSource)",
160                       data = na.omit(fecalsoil)[-1541,])
161
162 par(mfrow = c(1,3))
163 plot(cooks.distance(eco_model_385), type = 'h', lwd = 3, ylab = 'Cooks
164   ↪ distance',
165       main = 'Cooks distance Plot for Model w/o 385')
166 identify(1:nrow(model.matrix(eco_model_385)), cooks.distance(eco_model_385),
167          tolerance = 0.5, atpen = TRUE)
168
169 plot(cooks.distance(eco_model_137), type = 'h', lwd = 3, ylab = 'Cooks
170   ↪ distance',
171       main = 'Cooks distance Plot for Model w/o 137')
172 identify(1:nrow(model.matrix(eco_model_137)), cooks.distance(eco_model_137),
173          tolerance = 0.5, atpen = TRUE)
174
175 plot(cooks.distance(eco_model_1541), type = 'h', lwd = 3, ylab = 'Cooks
176   ↪ distance',
177       main = 'Cooks distance Plot for Model w/o 1541')
178 identify(1:nrow(model.matrix(eco_model_1541)), cooks.distance(eco_model_1541),
179          tolerance = 0.5, atpen = TRUE)
180
181 par(mfrow = c(1,1))
182 #####
183 #
184   ↪ #
185 #
186   ↪ #
187 #
188   ↪ #
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000

```

```

182  #
183    ↪  #
184  #####
185
186
187  summary(glm("Campylobact~factor(SampleType)+factor(Month)",
188             data = broilers[broilers$Farm == "A",], family =
189             ↪ binomial(link = "logit")))
190
191
192  #####
193  #
194    ↪  #
195  #
196    ↪  #
197  #
198    ↪  #
199  #####
200
201  attach(Compositional)
202  Sample_dummy <- factor(Compositional$Sampletype, labels = c(0,1))
203
204  summary(glm(Sample_dummy~A + B + C + D, family = binomial))
205  summary(glm(Sample_dummy~A + C + D + E, family = binomial))
206  summary(glm(Sample_dummy~A + B + D + E, family = binomial))
207  summary(glm(Sample_dummy~A + B + C + E, family = binomial))
208
209  compositional_model <- glm(Sample_dummy~A+ B + C + D, family =
210    ↪ binomial)
211  summary(compositional_model_A)
212
213  compositional_model_E <- glm(Sample_dummy~ B+ C+D+E, family = binomial)
214  summary(compositional_model_E)
215
216  #Predicted values
217  Pred_sample <- compositional_model_A$fit > 0.5

```

```

218 #Confusion matrix
219 table(Pred_sample, Samplettype)
220
221 ##Cross-Validation
222 test_num <- as.integer(nrow(Compositional)/3)
223 index <- sample(nrow(Compositional), test_num)
224 Sample_dummy_cross_val <- factor(Samplettype[-index], labels = c(0,1))
225 compos_cross_val_model <- glm("factor(Samplettype, labels = c(0,1)) ~
  ↪ A+B+C+D",
226                               data = Compositional[-index,], family =
  ↪ binomial)
227 Pred_sample_cross_val <- predict(compos_cross_val_model,
  ↪ Compositional[index, c("A", "B", "C", "D", "E")],
228                               type = "response") > 0.5
229 table(Pred_sample_cross_val, Samplettype[index])
230
231
232 ##Cross-Validation multiple loops
233 accuracy_score_list <- c()
234
235 for (i in seq(1,10)){
236   test_num <- as.integer(nrow(Compositional)/3)
237   index <- sample(nrow(Compositional), test_num)
238   Sample_dummy_cross_val <- factor(Samplettype[-index], labels = c(0,1))
239   compos_cross_val_model <- glm("factor(Samplettype, labels = c(0,1)) ~
  ↪ A+B+C+D",
240                               data = Compositional[-index,], family =
  ↪ binomial)
241   Pred_sample_cross_val <- predict(compos_cross_val_model,
  ↪ Compositional[index, c("A", "B", "C", "D", "E")],
242                               type = "response") > 0.5
243   correct_obs <-
  ↪ sum(diag(table(Pred_sample_cross_val, Samplettype[index])))
244   accuracy_score_list <-
  ↪ append(accuracy_score_list, (correct_obs/test_num))
245
246
247 }
248
249 summary(1-accuracy_score_list)
250
251 1-c(1,2,3)
252 predict(compos_cross_val_model,
  ↪ Compositional[index, c("A", "B", "C", "D", "E")],
253           type = "response")

```

```
254 detach(Compositional)
255
```