STAT6420
Final Project Cover Page December 5, 2023

Name : Agbolade Akande Student ID# 811811172

Pledge:

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

Signature:

Contents

List of Figures

List of Tables

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:

1 Conclusions

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%.

2 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?

3 Exploratory Data Analyses

3.1 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

$$\begin{aligned} \text{Camplyobacter} &= \beta_0 + \beta_1 \cdot \text{Salmonella} + \beta_2 \cdot \text{Listeria} \\ \text{Salmonella} &= \beta_0 + \beta_1 \cdot \text{Camplyobacter} + \beta_2 \cdot \text{Listeria} \\ \text{Listeria} &= \beta_0 + \beta_1 \cdot \text{Salmonella} + \beta_2 \cdot \text{Camplyobacter} \end{aligned}$$

The summary of these models are shown below:

```
> summary(model_Camplyobact)

Call:
lm(formula = "Campylobact˜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  1.000000000 -0.001109826  0.02057247
Listeria    -0.001109826  1.000000000 -0.02177444
Salmonella   0.020572470 -0.021774441  1.00000000
```

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

3.2 Second research question

We will study the effect on the sample type and the months on the presence/absence of *Camplyobacter*. 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:

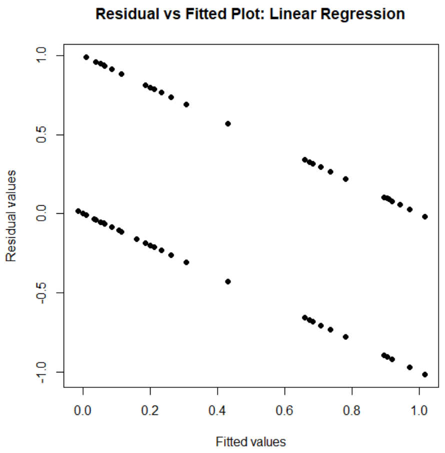

Figure 1: Residual vs fitted plot of regular linear regression

```
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.

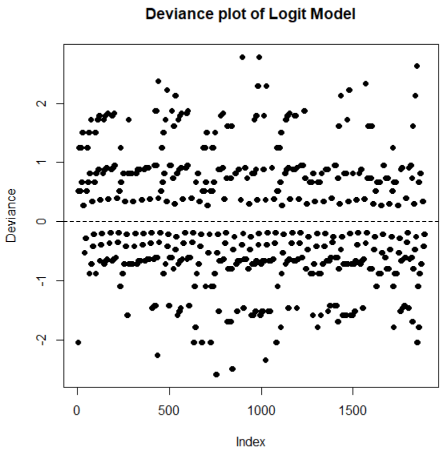

Figure 2: Deviance plot of logistic regression

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

3.3 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
(Intercept)                                        6.63741    0.05691
factor(SampleType)Soil                            -2.11200    0.08061
factor(AnimalSource)Cattle                        -1.68056    0.22025
factor(AnimalSource)Layer                          0.59819    0.16086
factor(AnimalSource)Swine                         -0.81281    0.20972
factor(SampleType)Soil:factor(AnimalSource)Cattle -0.90452    0.31151
factor(SampleType)Soil:factor(AnimalSource)Layer  -1.26201    0.22753
factor(SampleType)Soil:factor(AnimalSource)Swine  -2.41290    0.29662
                                                  t value Pr(>|t|)
(Intercept)                                       116.629  < 2e-16 ***
factor(SampleType)Soil                            -26.200  < 2e-16 ***
factor(AnimalSource)Cattle                         -7.630 3.98e-14 ***
factor(AnimalSource)Layer                           3.719 0.000207 ***
factor(AnimalSource)Swine                          -3.876 0.000111 ***
factor(SampleType)Soil:factor(AnimalSource)Cattle  -2.904 0.003739 **
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.

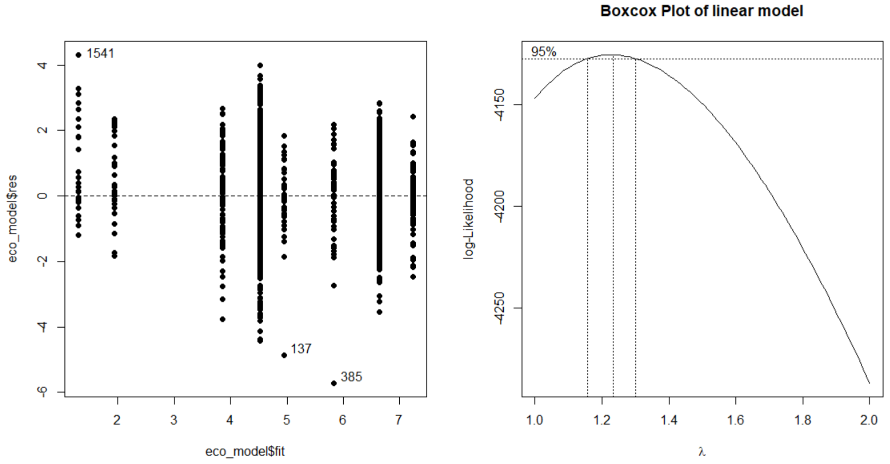

Figure 3: Residual vs Fitted Plot and Boxcox Plot of Linear regression

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

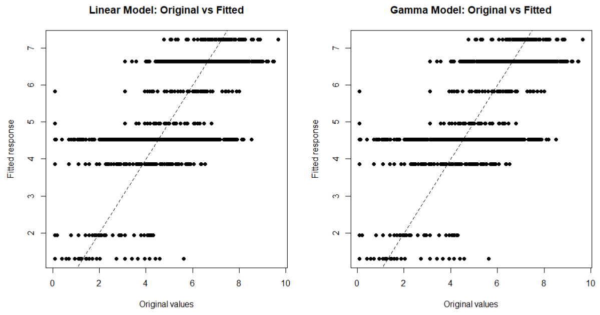

Figure 4: Fitted response vs original response for both linear regression fit and gamma regression fit

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

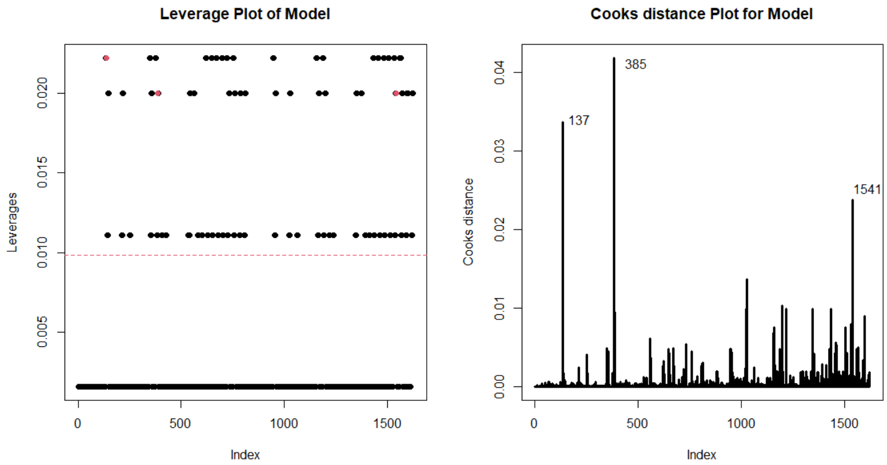

Figure 5: Leverage Plot and Cooks Distance Plot. The red dots are the entries 137, 385 and 1541.

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.

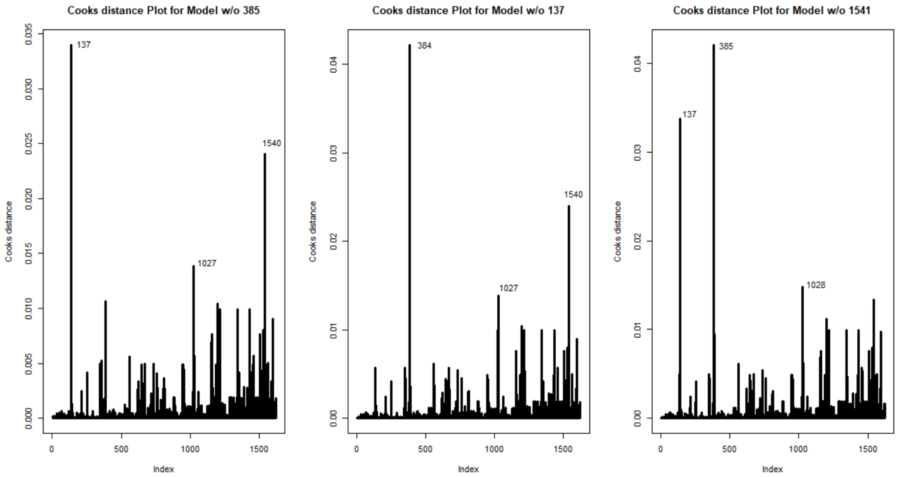

Figure 6: Cooks distances when the exceptional entries are removed

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.

3.4 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
```

4 Statistical Analyses

4.1 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.

4.2 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.

4.3 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 `Camplyobacter` 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.

4.4 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%.

5 Conclusions

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%.

6 Appendix

```
library(readxl)
library(faraway)
library(MASS)
library(leaps)

dev.new()

broilers <- read_excel("PasturedPoultryFarms.xlsx",sheet = "Broilers")

fecalsoil <- read_excel("PasturedPoultryFarms.xlsx",sheet =
 ↪  "FecalSoil")

Compositional <- read_excel("PasturedPoultryFarms.xlsx",sheet =
 ↪  "Compositional")


#######################################################################
  #
   ↪  #
  #
   ↪  #
  #                               First Ques
   ↪  #
  #
   ↪  #
  #
   ↪  #
  #######################################################################

model_Camplyobact <- lm("Campylobact~Listeria+Salmonella",data =
 ↪  broilers)
model_Salmonella <- lm("Salmonella~Campylobact+Listeria",data =
 ↪  broilers)
model_Listeria <- lm("Listeria~Salmonella+Campylobact",data = broilers)

summary(model_Camplyobact)
summary(model_Salmonella)
summary(model_Listeria)

model_Camplyobact_binom <- glm("Campylobact~Listeria+Salmonella",data =
 ↪  broilers,
                         family = binomial)
model_Salmonella_binom <- glm("Salmonella~Campylobact+Listeria",data =
 ↪  broilers,
```

```r
                              family = binomial)
model_Listeria_binom <- glm("Listeria~Salmonella+Campylobact",data = broilers,
                            family = binomial)

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


summary(model_Camplyobact_binom)
summary(model_Salmonella_binom)
summary(model_Listeria_binom)




#######################################################################
  #                                                                     #
  #                                                                     #
  #                                Second Ques                          #
  #                                                                     #
  #                                                                     #
  #######################################################################


campyl_model <- lm("Campylobact~factor(SampleType)+factor(Month)",
                   data = broilers)
summary(campyl_model)
plot(campyl_model$fit, campyl_model$res,pch = 19,xlab = "Fitted values",
     ylab = "Residual values",main = "Residual vs Fitted Plot: Linear Regression")

boxcox(lm("I(Campylobact+0.1)~factor(SampleType)+factor(Month)",
          data = broilers))

campyl_model_1 <- lm("log(Campylobact+0.1)~factor(SampleType)+factor(Month)",
                     data = broilers)
summary(campyl_model_1)
```

```

plot(campyl_model_1$fit, campyl_model_1$res,pch = 19)


campyl_model_binom_logit <- glm("Campylobact~factor(SampleType)+factor(Month)",
                              data = broilers, family = binomial(link = "logit"))
summary(campyl_model_binom_logit)


plot(residuals(campyl_model_binom_logit, type = "deviance"), pch = 19,
     ylab="Deviance",main ="Deviance plot of Logit Model")
abline(h = 0, lty = 2)

#Dispersion factor
campyl_model_binom_logit$deviance/campyl_model_binom_logit$df.residual

#######################################################################
  #                                                                     #
  #                                                                     #
  #                               Third Ques                            #
  #                                                                     #
  #                                                                     #
  #######################################################################



fecalsoil$EcoliLog10 <- as.numeric(fecalsoil$EcoliLog10)

eco_model <- lm("I(EcoliLog10 + 0.1) ~ factor(SampleType)+ factor(AnimalSource) + factor(SampleType)*factor(AnimalSource)",
                data = fecalsoil)
summary(eco_model)


par(mfrow = c(1,2))
plot(eco_model$fit,eco_model$res,pch = 19)
abline(h=0,lty = 2)
identify(eco_model$fit,eco_model$res,atpen = T, tolerance = 0.5)
```

```r

boxcox(eco_model, lambda = seq(1,2,by = 0.05))
title("Boxcox Plot of linear model")
par(mfrow = c(1,1))

step(eco_model)

Cpplot(leaps(model.matrix(eco_model)[,-1],na.omit(fecalsoil$EcoliLog10)+0.1))

maxadjr(leaps(model.matrix(eco_model)[,-1],na.omit(fecalsoil$EcoliLog10)+0.1,
              method="adjr2"), best = 8)

eco_model_gamma <- glm("I(EcoliLog10+0.1) ~ factor(SampleType)+ factor(AnimalSource) + factor(SampleType)*factor(AnimalSource)",
                       data = fecalsoil, family = Gamma(link = "inverse"))

summary(eco_model_gamma)


par(mfrow = c(1,2))
plot(na.omit(fecalsoil$EcoliLog10)+0.1,eco_model$fit,pch = 19,
     xlab = "Original values",ylab="Fitted response",main = "Linear Model: Original vs Fitted")
abline(a=0,b=1,lty = 2)

plot(na.omit(fecalsoil$EcoliLog10)+0.1,eco_model_gamma$fit,pch = 19,
     xlab = "Original values",ylab="Fitted response",main = "Gamma Model: Original vs Fitted")
abline(a=0,b=1,lty = 2)
par(mfrow = c(1,1))

par(mfrow = c(1,2))
plot(hat(model.matrix(eco_model)),pch = 19,ylab = 'Leverages',
     main = 'Leverage Plot of Model')
abline(h = 2*ncol(model.matrix(eco_model))/nrow(model.matrix(eco_model)),lty = 2, col = 2)
points(137,hat(model.matrix(eco_model))[137], col = 2,pch = 19)
points(385,hat(model.matrix(eco_model))[385], col = 2,pch = 19)
points(1541,hat(model.matrix(eco_model))[1541], col = 2,pch = 19)


plot(cooks.distance(eco_model),type = 'h',lwd = 3,ylab = 'Cooks distance',
     main = 'Cooks distance Plot for Model')
```

```r
identify(1:nrow(model.matrix(eco_model)),cooks.distance(eco_model),
         tolerance = 0.5, atpen = TRUE)
par(mfrow = c(1,1))

eco_model_385 <- lm("I(EcoliLog10 + 0.1) ~ factor(SampleType)+ factor(AnimalSource) + factor(SampleType)*factor(AnimalSource)",
                     data = na.omit(fecalsoil)[-385,])

eco_model_137 <- lm("I(EcoliLog10 + 0.1) ~ factor(SampleType)+ factor(AnimalSource) + factor(SampleType)*factor(AnimalSource)",
                         data = na.omit(fecalsoil)[-137,])

eco_model_1541 <- lm("I(EcoliLog10 + 0.1) ~ factor(SampleType)+ factor(AnimalSource) + factor(SampleType)*factor(AnimalSource)",
                         data = na.omit(fecalsoil)[-1541,])


par(mfrow = c(1,3))
plot(cooks.distance(eco_model_385),type = 'h',lwd = 3,ylab = 'Cooks distance',
     main = 'Cooks distance Plot for Model w/o 385')
identify(1:nrow(model.matrix(eco_model_385)),cooks.distance(eco_model_385),
         tolerance = 0.5, atpen = TRUE)

plot(cooks.distance(eco_model_137),type = 'h',lwd = 3,ylab = 'Cooks distance',
     main = 'Cooks distance Plot for Model w/o 137')
identify(1:nrow(model.matrix(eco_model_137)),cooks.distance(eco_model_137),
         tolerance = 0.5, atpen = TRUE)

plot(cooks.distance(eco_model_1541),type = 'h',lwd = 3,ylab = 'Cooks distance',
     main = 'Cooks distance Plot for Model w/o 1541')
identify(1:nrow(model.matrix(eco_model_1541)),cooks.distance(eco_model_1541),
         tolerance = 0.5, atpen = TRUE)

par(mfrow = c(1,1))

#######################################################################
  #                                                                       #
  #                                                                       #
  #                              Fourth Ques                              #
  #
   ↪  #
  #
   ↪  #
  #######################################################################


summary(glm("Campylobact~factor(SampleType)+factor(Month)",
            data = broilers[broilers$Farm == "A",], family =
             ↪  binomial(link = "logit")))



#######################################################################
  #
   ↪  #
  #
   ↪  #
  #                              Fifth Ques
   ↪  #
  #
   ↪  #
  #
   ↪  #
  #######################################################################


attach(Compositional)
Sample_dummy <- factor(Compositional$Sampletype, labels = c(0,1))

summary(glm(Sample_dummy~A + B + C + D, family = binomial))
summary(glm(Sample_dummy~A + C + D + E, family = binomial))
summary(glm(Sample_dummy~A + B + D + E, family = binomial))
summary(glm(Sample_dummy~A + B + C + E, family = binomial))

compositional_model <- glm(Sample_dummy~A+ B + C + D, family =
 ↪  binomial)
summary(compositional_model_A)

compositional_model_E <- glm(Sample_dummy~ B+ C+D+E, family = binomial)
summary(compositional_model_E)

#Predicted values
Pred_sample <- compositional_model_A$fit > 0.5

```

```
#Confusion matrix
table(Pred_sample, Sampletype)

##Cross-Validation
test_num <- as.integer(nrow(Compositional)/3)
index <- sample(nrow(Compositional), test_num)
Sample_dummy_cross_val <- factor(Sampletype[-index],labels = c(0,1))
compos_cross_val_model <- glm("factor(Sampletype,labels = c(0,1)) ~
 ↪  A+B+C+D",
                              data = Compositional[-index,], family =
                               ↪  binomial)
Pred_sample_cross_val <- predict(compos_cross_val_model,
 ↪  Compositional[index,c("A","B","C","D","E")],
                                 type = "response")>0.5
table(Pred_sample_cross_val,Sampletype[index])


##Cross-Validation multiple loops
accuracy_score_list <- c()

for (i in seq(1,10)){
  test_num <- as.integer(nrow(Compositional)/3)
  index <- sample(nrow(Compositional), test_num)
  Sample_dummy_cross_val <- factor(Sampletype[-index],labels = c(0,1))
  compos_cross_val_model <- glm("factor(Sampletype,labels = c(0,1)) ~
   ↪  A+B+C+D",
                                data = Compositional[-index,], family =
                                 ↪  binomial)
  Pred_sample_cross_val <- predict(compos_cross_val_model,
   ↪  Compositional[index,c("A","B","C","D","E")],
                                   type = "response")>0.5
  correct_obs <-
   ↪  sum(diag(table(Pred_sample_cross_val,Sampletype[index])))
  accuracy_score_list <-
   ↪  append(accuracy_score_list,(correct_obs/test_num))


}

summary(1-accuracy_score_list)

1-c(1,2,3)
predict(compos_cross_val_model,
 ↪  Compositional[index,c("A","B","C","D","E")],
        type = "response")
```

```
detach(Compositional)
```