STA 207 Final Project

David Huang, Kai Jin, & Niklas Braun

March 23, 2017

Abstract

A study of the survival rates of swallows after a sever storm in the winter of 1898 recorded the number of birds found on the ground after the storm subsided, and whether or not the birds survived the storm. In modeling the data in attempt to predict the survival of the birds based on physical characteristics, a step-wise selection was used to select an appropriate logistic regression model. The resulting model relies on four physical aspects of the birds in question. However, the model is hindered in effectiveness by the small amount of data in the dataset.

1 Introduction

In this report we look at the occurrences of swallow deaths that occurred during and after a 1898 winter storm. Many swallows were found on the ground in the aftermath. Some of the birds died, while others survived. The goal of this study is to determine whether a relationship exists between the probability that a bird survived and its physical characteristics.

The data for this study contains ten independent variables and one response variable. The response variable is a binomial response, whether or not the swallow survived. The ten independent variables are age (juvenile or adult), total length of the swallow, Alar extent (tip-to-tip wing length), weight, length of back and head, length of humerus, length of femur, length of tibio-tarsus, width of skull, and length of keel of sternum.

As can be seen, all of the variables are physical characteristics of the birds. It is hoped that some or all of these characteristics can be used to predict whether or not a bird will have survived the storm.

2 Methodology

2.1 Exploratory Analysis

The non-binomial variables are examined with histograms of every plot (shown in Figure 3). Four of the variables are not normally distributed; for these, a log-transformation is used. The data is then standardized. The fixed data has marginally better distributions (shown in Figure 4).

It is also important to obtain the distribution of values of the response. Since it is bimodal, there are only two responses (dead or alive). These are shown here in figure 1.

Correlation between variables is also investigated for all variables. This is important not only because there are a large number of similar variables, but also because it will help decide what type of regression will be used. A simple correlation plot shows a moderate amount of correlation between many of the variables (Figure 6). This can be taken a step further to show the clusters of variable groups as represented in their respective response variable values (Figure 5). This figure also contains boxplots for each variable. The lone binomial variable is checked for distribution between response values as well (Figure 7).

The histograms of the variables reveal that many of the variables may be unnecessary in the fitting of the model. There is a large number of correlated variables, and many don't appear to provide insight to whether the swallow will live or die. Some variables, such as TL (total length) will certainly be valuable in prediction according to the correlation plot.

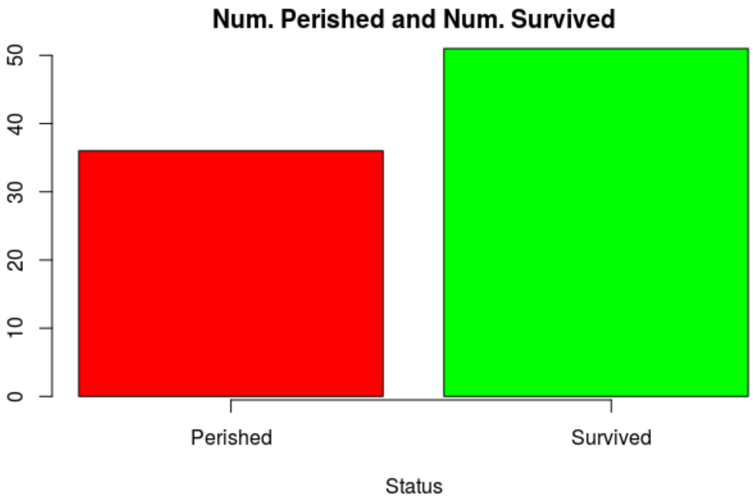

Figure 1: Swallow Outcomes after Storm

2.2 Statistical Modeling

Data validation requires a test dataset and a training dataset. Unfortunately, only one dataset of this type exists; however, it is possible to split the data that is available into two to mimic testing and training datasets. That is what is done here; the training dataset is used to fit the variables, and validated with the testing dataset. These are split at 80%-20%, meaning the training dataset contains 4/5 of the data, and the testing dataset contains the rest. This large split is required because the dataset itself is small, so to maximize accuracy most of the data is dedicated to training the model.

To fit this data, a method that predicts a binomial output must be used. To get the most accurate model possible, two types of binomial regression are used, then compared for accuracy. The type of modeling that will be used is a regression model, which is selected in a step-wise procedure. The full models considered for selection are a first order model and a second order model with interactions.

2.2.1 Stepwise Logistic Regression

The first method used to fit the data is a cross-validation logistic regression. The model to use is selected via a stepwise regression. This starts with a full model, then excludes variables until a certain criteria is minimized. The criterion minimized included AIC, BIC, and MSPE. The full models considered included one without interaction terms, and one with both second order terms and a number of two-way interaction terms. It was not possible to include a model with all two-way interactions because of number of variables, which proved too many for the computational abilities of R on the test computers. Because of this, intuition was used to select interaction terms which appeared to have the most collinearity. In this case, only the interaction between total length and weight is considered.Note that total length and weight also have the largest correlation amount all the other 4 parameters we got from first order stepwise in Table2.

Four models were generated as a result of these: two models of the non-interaction terms (minimized by AIC and BIC), and two models containing interaction and second order terms (again, minimized by AIC and BIC). The residual vs. fitted value plots as well as the normal probability plots are generated for each model. These values are contained in Table 2.

After generating these models, they are validated with the validation dataset. Finally, by finding the mean-squared prediction error for each of these models, the best model is selected. The four model selection procedures actually resulted in only two unique models; one model was generated by three of the algorithms.

2.2.2 Residual Diagnostics

After finding a preliminary model, diagnostics are run on it to test its goodness of fit. The residuals are plotted and compared to the fitted model. The Q-Q plots are also examined. Outliers are also investigated, an important step considering how small the sample size is.

3 Results

3.1 Stepwise Regression

The best model selected in stepwise regression is

$$y_{Status} = \beta_0 + \beta_1 x_{TL} + \beta_2 x_{WT} + \beta_3 x_{HL} + \beta_4 x_{KL}$$

where the coefficients β_0, β_1, β_2, β_3, and β_4 are given in Table 1, along with the standard errors and the P-value of the test statistic for each coefficient.

Parameter	Estimate	Std. Error	Z Value	P-value
β_0	0.8221	0.3563	2.307	0.02103
β_1	-1.8936	0.5709	-3.317	0.00091
β_2	-1.0641	0.4775	-2.229	0.02584
β_3	1.4510	0.4946	2.933	0.00335
β_4	0.9108	0.4324	2.107	0.03515

Table 1: Stepwise Model ANOVA.

This was selected by minimizing AIC and BIC for each respective full model, as stated earlier. The iterative process for these minimizations can be seen in figure 8

The model is tested to make sure that the reduced model is in fact representative of the data. This is done by comparing the difference of the residual deviations from the to models to a χ^2 distribution:

$$Res.Dev.(F) - Res.Dev.(R) = 57.415 - 59.452 < \chi^2(0.95, 1) = 3.841$$

Because it is lower, we can safely exclude the extra term from model 2 and conclude that model 1 is sufficient in modeling the data.

3.2 Residual Diagnostics

In the Figure 9 and Figure 10, the residuals plots lines on the fitted line, a function of the predicted value. Residuals show both models are in a good fit. Notice that in the full model there may be an outlier in the residual plots. The reduced model does not have outliers according to the residual plots. The QQ plots of reduced model is followed closely on the line and the QQ plots have a little bit skew which is not as good as the reduced model.

Residual plots show that the reduced model is a better fit of the data, confirming what was found in the previous section. However, considering the nature of the data, it is important to also consider over-fitting. Therefore potential outliers must be explored. This is done in figure 11, where the Cook's distance of the reduced model is checked and compared with the cut-off of value $2 * \#Var/n = 0.145$. No value exceeds this threshold, so no outliers need be removed.

3.3 Goodness of Fit

Now the model may be tested on the test data. As mentioned earlier, the test data is 20% of the given total data. The model is imposed upon this, and if the result for a sparrow is that it has a greater-than-50% chance of survival, it is marked as alive; and conversely, if it is less than 50%, then it is marked as perished. This is compared to the actual status of the animal. The result is shown below in figure 2.

The result is that 15/18, or 83%, of the outcome statuses of the sparrows were correctly predicted by the model. The model appears to predict the data at a reasonable level. Though this is not a perfect model, it is most likely the best model that could be made considering the small amount of input data for both fitting and testing that is available.

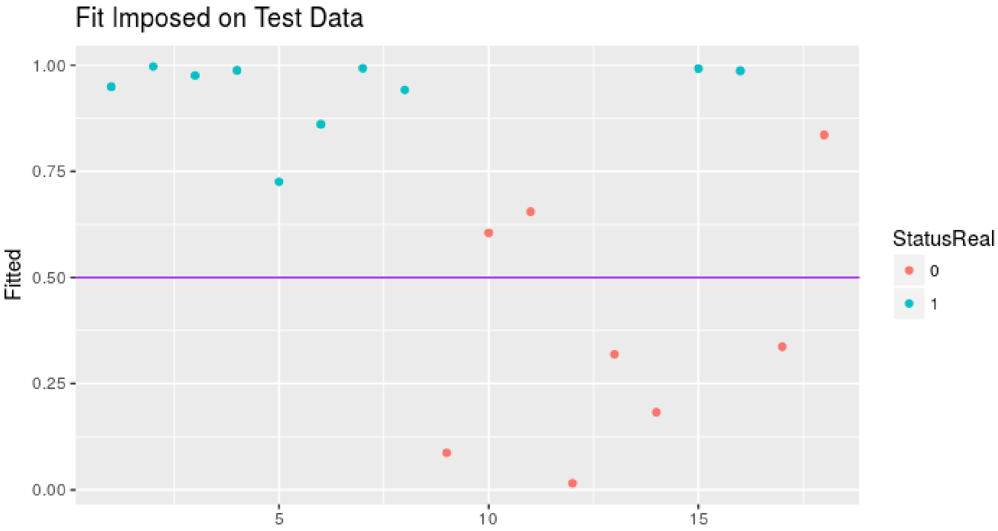

Figure 2: Model Testing, 20% of Original Data

4 Discussion

The stepwise regression model provided an acceptable model for the dataset. It has around an 80% accuracy rate, which is not terrible considering the small nature of the dataset. This could be greatly increased if another dataset, or a larger dataset, were presented.

The result shows that keep everything else constant, every one unit increase in total length, the odds of survival changes by $exp(-1.8936) = 0.15$, or it decreases by 85%. For every one unit increase in weight, the odds of survival changes by $exp(-1.0641) = 0.35$, or decreases by 65%. For every one unit increase in the length of the humerus, the odds of survival is $exp(1.451) = 4.26738$ times greater; and finally, for every one unit increase in the length of the keel of the sternum, the odds of survival is $exp(.9108) = 2.486311$ times greater.

What this means is that sparrows of smaller overall size, but with a large humerus and sternum, were more likely to survive the storm. Intuitively, this could be because smaller birds we able to get out of the brunt of the storm more easily, though this is pure speculation.

5 Conclusion

A step-wise selected regression model fit the sparrow death data adequately considering the limits of the dataset. Of the many variables, only four were shown to have significance to the data at a 95% confidence level, after weeding out other variables through AIC and BIC criteria. This was so even after accounting for interaction and second order terms. The model was also confirmed when testing for residual deviance between it and other models, as well as outliers.

The selected model shown in table 2 is shown to have an 83% accuracy when predicting the outcome status on the test data. Though not great, it is good considering the fact that the dataset provided contains a small number of data points, hindering its usefullness. In the future, more data would help increase the accuracy of prediction in the model for the sparrows.

6 Appendix

6.1 Appendix A: Data Exploration

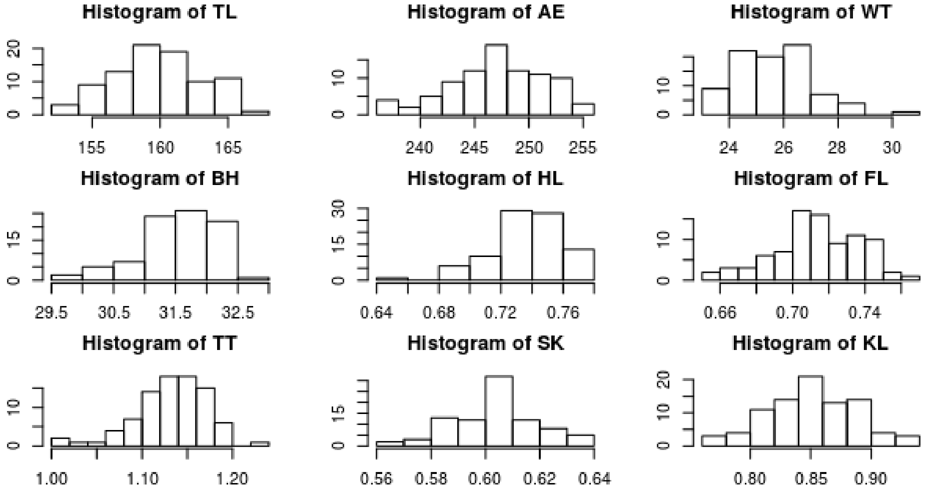

Figure 3: Histogram of Raw Data Variables (Non-Binomial)

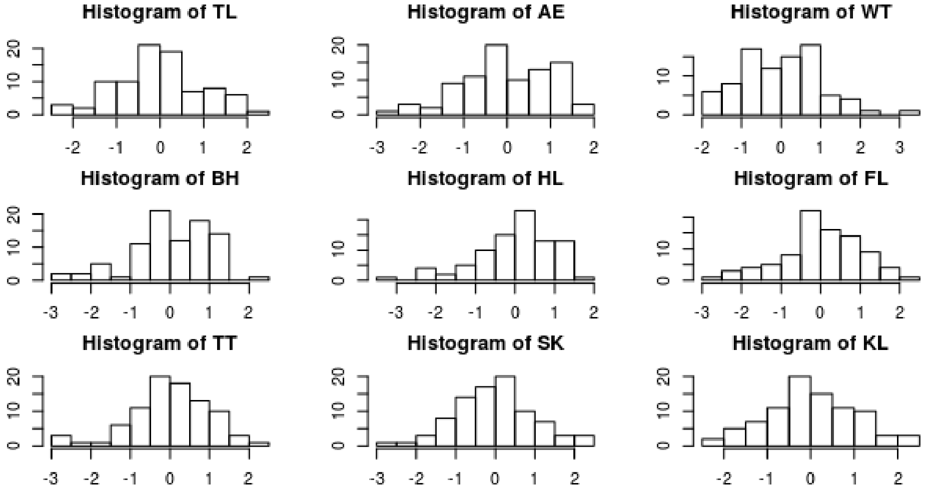

Figure 4: Histogram of Transformed & Standardized Data Variables (Non-Binomial)

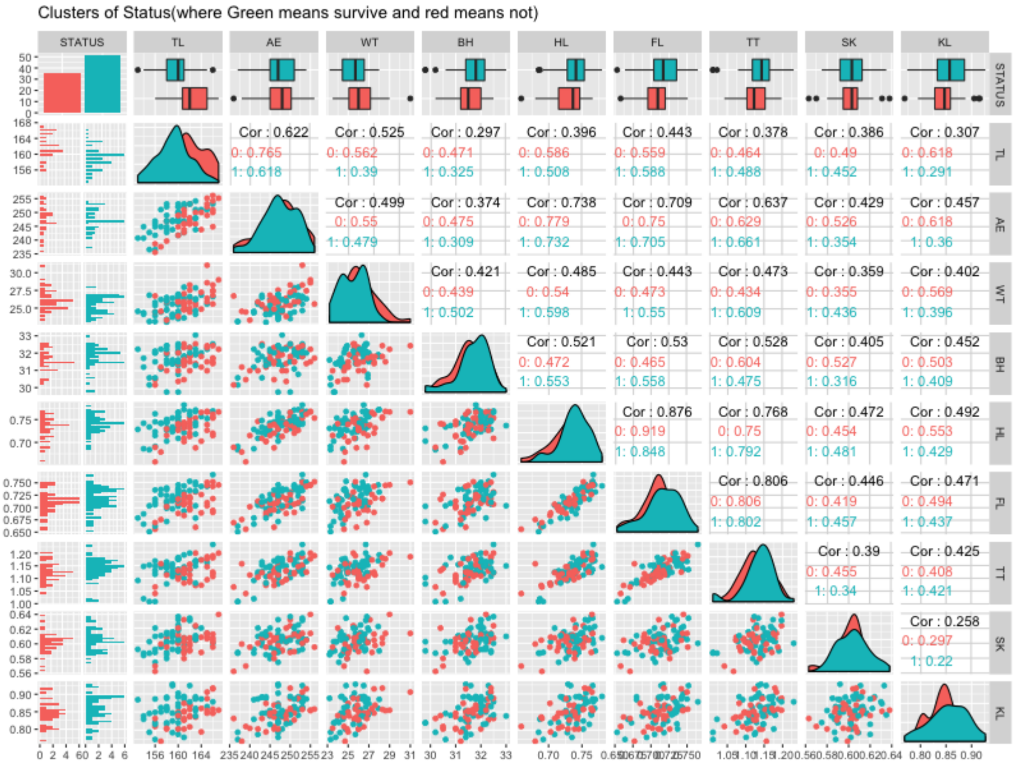

Figure 5: Clusters and Distributions of Variables with Respect to Status (Response)

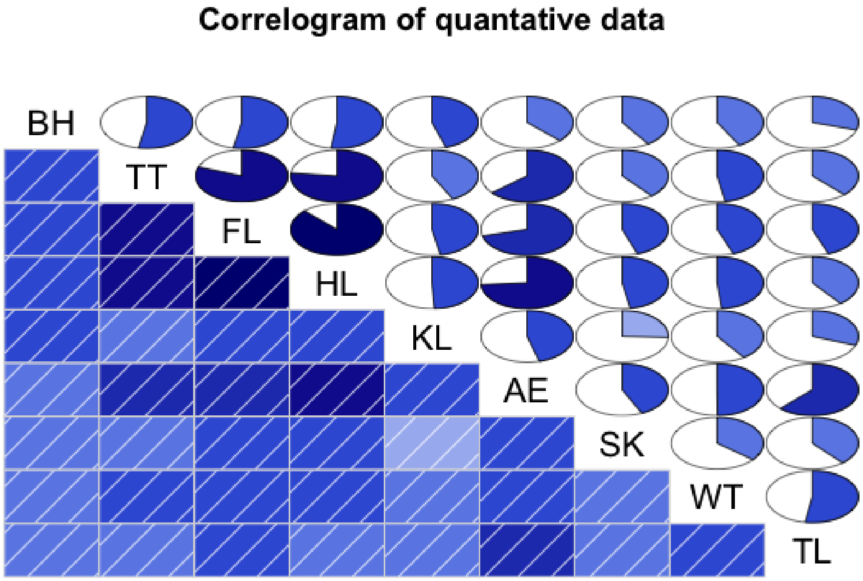

Figure 6: Correlation of Quantitative Variables

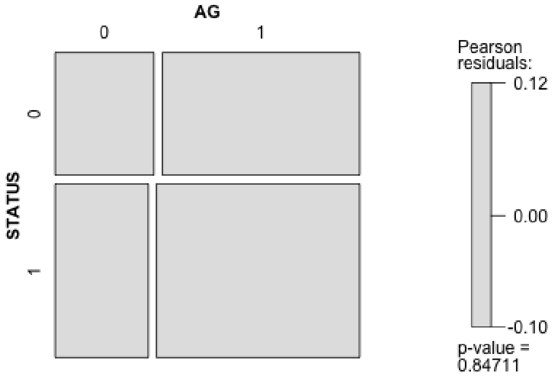

Figure 7: Categorial Variables

6.2 Appendix B: Model Selection

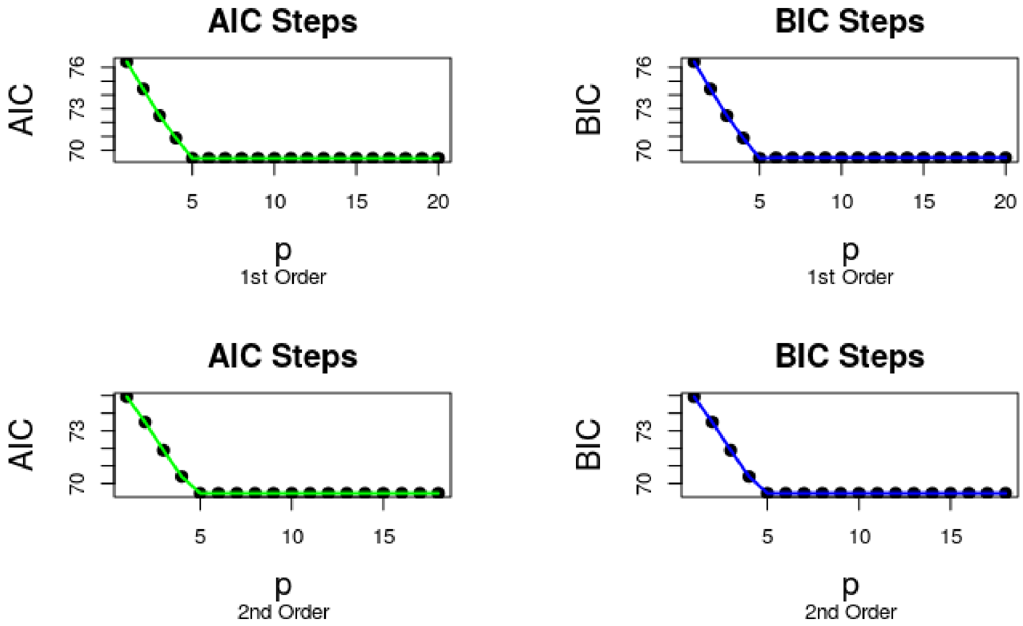

Figure 8: Iterative AIC and BIC

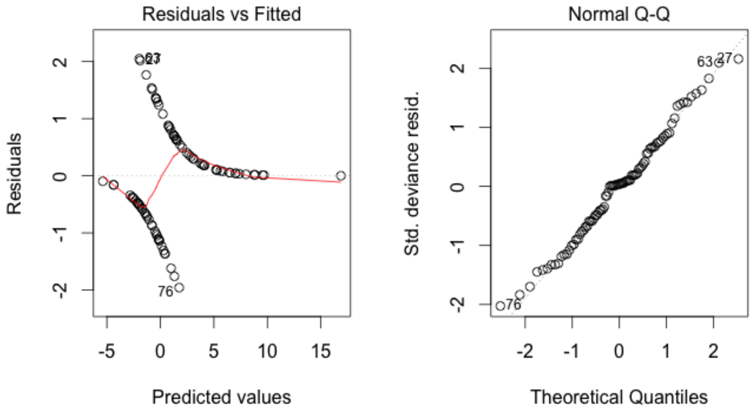

Figure 9: residual and QQ plots for full model

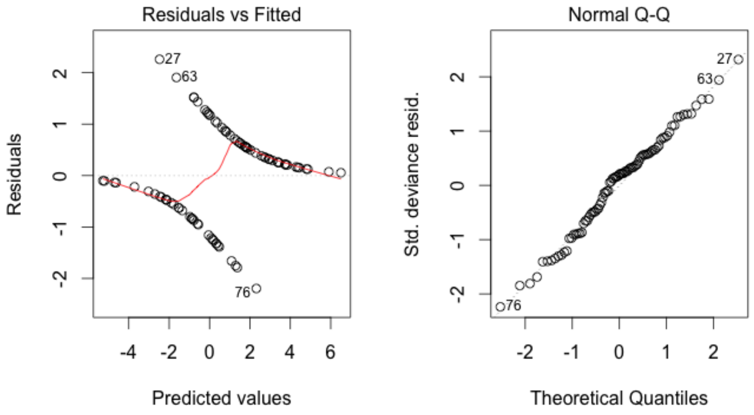

Figure 10: residual and QQ plots for reduced model

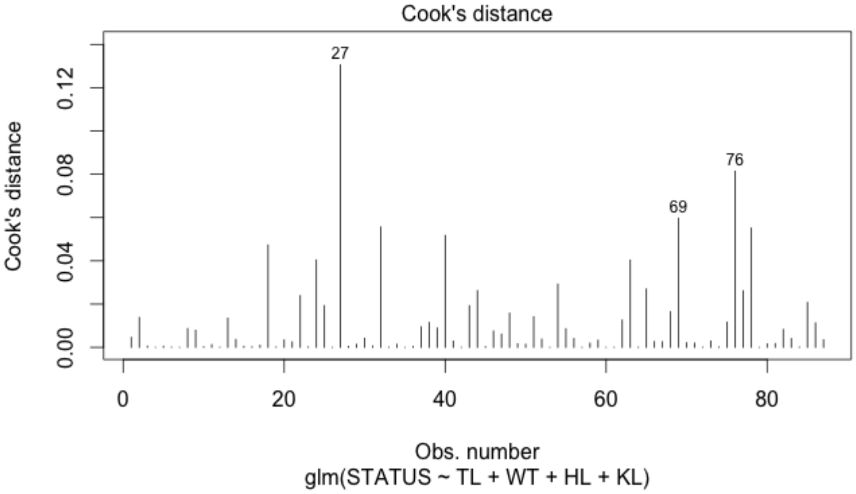

Figure 11: Cook distance

Type	Criteria	Included Terms in Model	MSPE
1^{st} Order	AIC = 69.4522	$\beta_0 + \beta_1 x_{TL} + \beta_2 x_{WT} + \beta_3 x_{HL} + \beta_4 x_{KL}$	0.103
1^{st} Order	BIC = 69.41539	$\beta_0 + \beta_1 x_{TL} + \beta_2 x_{WT} + \beta_3 x_{BH} + \beta_4 x_{HL} + \beta_5 x_{KL}$	0.18
2^{st} Order + Interaction	AIC = 69.4522	$\beta_0 + \beta_1 x_{TL} + \beta_2 x_{WT} + \beta_3 x_{HL} + \beta_4 x_{KL}$	0.110
2^{st} Order + Interaction	BIC = 69.4522	$\beta_0 + \beta_1 x_{TL} + \beta_2 x_{WT} + \beta_3 x_{HL} + \beta_4 x_{KL}$	0.103

Table 2: Stepwise Selection Final Models.

6.3 Appendix C: Code

```r
library(ggplot2)
library(MASS)
library(corrgram)
library(purrr)
library(GGally)
library(vcd)
# Useful functions ----
logit = function(x) {log(x / (1 - x))}
inv.logit = function(x) {exp(x)/(1 + exp(x))}

# Import and cleaning ----
survival <- data.frame(readxl::read_excel(path =
  '~/MEGAsync/School/Classes/STA/207/Project/survival_sparrow.xls',
  col_names = T))
survival$STATUS[survival$STATUS == 'Survived'] = 1
survival$STATUS[survival$STATUS == 'Perished'] = 0
survival$AG[survival$AG == 2] = 0
survival$STATUS <- factor(survival$STATUS)
survival$AG <- factor(survival$AG)

# Data transform and normalization ----
# First check distribution
par(mfrow = c(3, 3), mai = c(0.2, 0.2, 0.4, 0.4))
for (i in 3:11) {
  hist(survival[, i], xlab = names(survival)[i],
    main = paste("Histogram of", names(survival)[i]))
}

# Need to log transform WT, BH, HL, and FL
for (colz in c('WT', 'BH', 'HL', 'FL')) {
  survival[, colz] <- log(survival[, colz])
}

# Can also standardize all the variables
for (i in 3:11){
  survival[, i] <- (survival[, i] - mean(survival[, i]))
  survival[, i] <- survival[, i] / sd(survival[, i])
}

# Recheck Distribution
par(mfrow = c(3, 3), mai = c(0.2, 0.2, 0.4, 0.4))
for (i in 3:11) {
  hist(survival[, i], xlab = names(survival)[i],
    main = paste("Histogram of", names(survival)[i]))}

#########################
# exploratory data analysis
#########################

# summary of data
summary(survival)
```

```
# checking the correlation of vairables
corrgram(survival[,c(-1)], order = TRUE,
        lower.panel = panel.shade, upper.panel = panel.pie,
        text.panel = panel.txt,
        main = "Correlogram of quantative data")
par(mfrow = c(1,1))
plot(survival$STATUS, xlab = 'Status',
    main = 'Num. Perished and Num. Survived',
    col = c('red', 'green'), xaxt='n')
axis(1, c(.7, 1.95), c('Perished', 'Survived'))

# checking the clusters of quantative variables with respect to Status
ggpairs(data=survival[,c(-2)],
        mapping=ggplot2::aes(colour = STATUS),
        title ="Clusters of Status(where Green means survive and red means not)")

# checking the categorical variables
mosaic(~ STATUS + AG, data = survival, shade = TRUE, legend = TRUE)


# Initial full model fit ----
full.model <- glm(STATUS ~ ., data = survival, family = binomial())

# matrix scatter plot
pairs(survival)
# correlation matrix
corrplot::corrplot(survival[, -c(1, 2)])

# VIF Checking
(vifmat <- diag(solve(cor(survival[, -c(1:2)]))))

##########################################################
# cross validation
##########################################################

set.seed(100)
n.s=nrow(survival) ## number of cases in data
index.s=sample(1: n.s, size=n.s*.8, replace=FALSE)
## randomly sample 183 cases to form the training data.
survival.c=survival[index.s,]  ## get the training data set.
survival.v=survival[-index.s,]
n=nrow(survival.c)


# Initial full model fit ----
full.model <- glm(STATUS ~ ., data = survival.c, family = binomial())
null.model<- glm(STATUS ~ 1, data = survival.c, family = binomial())

# Stepwise AIC model fitting ----

# TL + WT + HL + KL BIC: 69.4154
# TL + WT + HL + KL AIC: 69.4522
Model1=stepAIC(full.model, direction = 'both', k=log(n))
```

```
Model2=stepAIC(full.model, direction = 'both', k=2)
getAICBIC <- function(full.model) {
  n <- length(full.model$residuals)
  num_terms <- length(full.model$coefficients) - 1
  aics <- c()
  bics <- c()
  thismod <- full.model
  thatmod <- full.model
  for (i in 1:(num_terms * 2)) {
    thismod <- stepAIC(thismod, direction = 'both', k = 2, steps = 1)
    thatmod <- stepAIC(thatmod, direction = 'both', k = log(n), steps = 1)
    aics <- append(aics, thismod$aic)
    bics <- append(bics, thatmod$aic)
  }

  return(list(AIC = aics, BIC = bics))
}

plotAICBIC <- function(AICBIC) {
  aics <- AICBIC$AIC
  bics <- AICBIC$BIC
  par(mfrow = c(1,2), mai = c(1, 1, 0.5, 0.5))
  plot(1:length(aics), aics, xlab="p", ylab="AIC",
    pch=19, cex=1.2, cex.axis=1, cex.lab=1.5, cex.main=1.5,
    main = 'AIC Steps')
  points(1:length(aics), aics, type='l', lwd=2, col = 'green')
  plot(1:length(bics), bics, xlab="p", ylab="BIC",
    pch=19, cex=1.2, cex.axis=1, cex.lab=1.5, cex.main=1.5,
    main = 'BIC Steps')
  points(1:length(bics), bics, type='l', lwd=2, col = 'blue')
}

# No Interaction Model ----
simple.model <- getAICBIC(full.model)
plotAICBIC(simple.model)

# +I(TL*HL) +I(TL*KL) +I(HL*WT) +I(KL*WT) +I(HL*KL)
full.model2 = glm(STATUS ~ TL + WT + HL + KL + I(TL^2)+I(WT^2)+I(HL^2)+I(KL^2)+I(TL*WT) , data = surviva

#  STATUS ~ TL + WT + HL + KL BIC: 69.4522
Model3=stepAIC(full.model2, direction = 'both', k=log(n))
# TL + WT + HL + KL + I(TL^2) + I(KL^2) AIC: 69.4522
Model4=stepAIC(full.model2, direction = 'both', k=2)

# Interaction Model ----
complex.model <- getAICBIC(full.model2)
plotAICBIC(complex.model)

par(mfrow=c(1,2))
plot(Model3,which=1)
plot(Model3,which=2)

plot(Model4,which=1)
plot(Model4,which=2)

plot(Model1,which=1)
plot(Model1,which=2)

plot(Model2,which=1)
plot(Model2,which=2)


###################
#validate the result
###################
m=nrow(survival.v)
head(survival.v)
newdata=survival.v[,-1]
# predict the values
inv.logit = function(x) {exp(x)/(1 + exp(x))}
ppi.prime1 <- predict.glm(Model1, newdata)
pred.fs1=inv.logit(ppi.prime1)

ppi.prime2 <- predict.glm(Model2, newdata)
pred.fs2=inv.logit(ppi.prime2)

ppi.prime3 <- predict.glm(Model3, newdata)
pred.fs3=inv.logit(ppi.prime3)

ppi.prime4 <- predict.glm(Model4, newdata)
pred.fs4=inv.logit(ppi.prime4)

# find the numeric value of survive
ValX=as.numeric(survival.v[,1])-1
# find MSPE
mspefs1=mean((pred.fs1-ValX)^2)
mspefs2 <- mean((pred.fs2-ValX)^2)
mspefs3 <- mean((pred.fs3-ValX)^2)
mspefs4=mean((pred.fs4-ValX)^2)

c(mspefs1,mspefs2, mspefs3, mspefs4)

## mspe for model 1 is the smallest
# so choose modelbs3: STATUS ~ TL + WT + FL + KL

# To test if the reduced model accurately models the data, we can do a Residual
# test comparing the two different models to see if the reduced model is indeed
# enough. This difference is compared to a chi-squared distribution at 95%:
# H0: G^2 < Chi-squared(0.95, 1)
# Ha: G^2 > Chi-squared(0.95, 1)
abs(Model2$deviance - Model1$deviance) < qchisq(0.95, 1)
# [1] TRUE
# Because it is less, H0 is concluded: that the reduced model is sufficient.

# Cook's Distance ----
par(mfrow = c(1,1))
```

```
plot(Model1, which = 4)

summary(Model3)

# Testing model fit
survival.vtest <- survival.v[, c('TL', 'WT', 'HL', 'KL')]
coefs <- Model1$coefficients
survival.vtest$Fitted <- inv.logit(predict.glm(Model1, survival.vtest))
survival.vtest$StatusFit[survival.vtest$Fitted >= .5] <-  1
survival.vtest$StatusFit[survival.vtest$Fitted < 0.5] <-  0
survival.vtest$StatusReal <- factor(survival.v$STATUS)
survival.vtest$StatusFit <- factor(survival.vtest$StatusFit)
survival.vtest

ggplot(survival.vtest, aes(x = 1:18, y = Fitted, color = StatusReal)) +
  geom_point() + geom_hline(yintercept = 0.5, color = 'purple') +
  xlab(label = "") + labs(title = 'Fit Imposed on Test Data')
```