Analysis of Sparrow Survival Using Physical Characteristics

Michael Chen
maichen@ucdavis.edu

Douglas Turner
dcturner@ucdavis.edu

March 17, 2018

1 Introduction

1.1 The Data

The data consists of 87 observations on house sparrows that were found injured after a severe winter storm hit Rhode Island on February 1, 1898. Of the 87 sparrows, 36 ended up perishing while 51 survived; this is the main variable of interest. In total, eleven different variables were recorded. They are summarized below:

- `STATUS`: Survived or perished.
- `AG`: Age, where adults are labeled as 1 and juveniles are labeled as 2.
- `TL`: Total length.
- `AE`: Alar extent.
- `WT`: Weight.
- `BH`: Beak and head length.
- `HL`: Humerus length.
- `FL`: Femur length.
- `TT`: Tibio-tarsus length.
- `SK`: Skull width.
- `KL`: Keel of sternum length.

As we can see, most of the variables describe various physical characteristics of the birds. Survival status and age are categorical (factor) variables, and the rest are continuous variables.

1.2 Aim of the Study

The goal of this study is to examine whether we can predict survival probability using some combination of the other variables in a logistic regression model. While it may be difficult to generalize the results due to the specific setting in which this data was collected, it is nevertheless instructive to see which characteristics were beneficial for the survival of these sparrows.

2 Preliminary Data Analysis

2.1 Data Visualization

In Figure 1, we see that many of the continuous variables are medium to highly correlated with each other. This is not surprising - as mentioned earlier, the variables measure physical characteristics, which tend to increase with age (at least up to a certain point). The scatterplot matrix of the continuous variables in Figure 2 indicate that there is a linear relationship between many of the predictors themselves.

Figure 3 provides a visual display of survival proportion by age group. Perhaps surprisingly, the proportion of survival between the two age groups is very similar: 53.9% in adults and 57.1% in juveniles. Figure 4 displays histograms of the univariate distributions of the nine continuous predictors after they have been standardized to have zero mean and unit variance. The variables are standardized because they measure in different unit scales (weight vs length). After standardization, we can more easily compare magnitudes of regression coefficients. We note that: (1) all the distributions appear to be unimodal; (2) total length, tibio-tarsus length, skull width, and keel of sternum length appear to be symmetric and possibly normally distributed; (3) alar extent, beak and head length, humerus length, and femur length appear to be left-skewed; and (4) weight appears to be right-skewed.

Next, we separated the data within each of the continuous predictors into bins of equal size and then plotted the proportion of survival within each bin, as can be seen in Figure 5. From this, it appears that total length, weight, and humerus length may be reasonably modeled using a logit relationship with survival proportion. After performing a logit transformation on the response, we obtain the plots in 6. Now, we see that total length, weight, and humerus length appear to have close to a monotone linear relationship with logit of the response, while the other predictors do not have a discernible relationship with logit of the response.

We conclude our exploratory graphical analysis by plotting the original binary response variable against each of the predictors and then fitting a smoothing line to each plot, which can be seen in Figure 7. Here, a flat line indicates little to no relationship between the predictor and the response, while a sloped line indicates that there may be a relationship. Once again, we can see that total length, weight, and humerus length appear to have a relationship with survival, while the rest are inconclusive.

2.2 Initial Model

We will use a logistic regression model with survival as the response variable and the other variables as the pool of potential predictors. Our initial model will contain all the predictors with no higher-order or interaction terms. Additionally, since the variables appear to be on different scales, we will standardize the continuous variables before proceeding.

We consider the responses Y_i conditional on $\mathbf{X}_i$ to be random and distributed as Bernoulli(π_i), where π_i is the probability of survival. Hence, $\mathbb{E}(Y_i|\mathbf{X}_i) = \pi_i$. The model is as follows:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1\text{juvenile}_i + \beta_2\text{TL}_i + \beta_3\text{AE}_i + \beta_4\text{WT}_i + \beta_5\text{BH}_i + \beta_6\text{HL}_i + \beta_7\text{FL}_i + \beta_8\text{TT}_i + \beta_9\text{SK}_i + \beta_{10}\text{KL}_i$$

Note that juvenile$_i$ is an indicator variable that takes the values of 1 (if the sparrow is juvenile) and 0 (if the sparrow is an adult). Equivalently, we can express this directly through π_i:

$$\pi_i = \frac{\exp(\beta_0 + \beta_1\text{juvenile}_i + \beta_2\text{TL}_i + \beta_3\text{AE}_i + \beta_4\text{WT}_i + \beta_5\text{BH}_i + \beta_6\text{HL}_i + \beta_7\text{FL}_i + \beta_8\text{TT}_i + \beta_9\text{SK}_i + \beta_{10}\text{KL}_i)}{1 + \exp(\beta_0 + \beta_1\text{juvenile}_i + \beta_2\text{TL}_i + \beta_3\text{AE}_i + \beta_4\text{WT}_i + \beta_5\text{BH}_i + \beta_6\text{HL}_i + \beta_7\text{FL}_i + \beta_8\text{TT}_i + \beta_9\text{SK}_i + \beta_{10}\text{KL}_i)}$$

After looking at the initial model, we will add all quadratic and interaction terms to the pool of potential predictors. Since we standardized all the variables, there should not be any major multicollinearity issue with including the quadratic terms. We will then run a forward stepwise regression using the AIC criterion to choose a final model for consideration.

2.3 Initial Analysis

As described in the previous subsection, our initial model is a logistic model with survival (binary) as the response and all first-order terms as the predictors. The output is displayed in Figure 8. As we can see, this initial analysis highlights total length and weight as significant predictors after accounting for the effect other the other covariates, while humerus length and keel of sternum length are almost significant. This matches our visual analysis from earlier, where we saw that total length, weight, and humerus length appeared to have a potential logistic relationship with survival. The variance inflation factor (VIF) scores indicate that there is some multicollinearity, which is consistent with what we saw in the correlation matrix of the continuous predictors in Figure 1.

Looking at the Analysis of Deviance tables comparing our initial model to the null and saturated models respectively (Figure 9), we see that the initial model is significantly better than the null model and not significantly worse than the saturated model. Thus, it appears to be a reasonable model for predicting the response, but perhaps we can do better. We will now proceed to model selection.

3 Model Selection and Diagnostics

3.1 Model Selection and Interpretation

As mentioned previously, we will use forward stepwise regression to choose from a pool of predictors that include all first-order terms, all quadratic terms, and all second-order interaction terms. Let us begin by analyzing the performance of the model that includes all first-order and quadratic terms.

In Figure 10, we see from the output that the only significant quadratic term (in the presence of all other linear and quadratic terms) is that of femur length. The Analysis of Deviance table indicates that this quadratic model is not significantly superior to our initial model, but it is close (p-value = 0.1069). In Figure 11, we see that once again, there is an issue of multicollinearity in this set of predictors. Now, let us proceed to run the stepwise procedure.

Figure 12 shows the result of the last step performed in the stepwise procedure (the intermediate steps have been redacted due to their length). We see that the final model chosen using the AIC criterion includes total length, humerus length, weight, keel of sternum length, beak and head length, and the square of femur length as predictors. However, this model (the ”preliminary final model”)is potentially problematic since it includes the square of femur length variable but not the linear femur length variable. To decide whether we want to include femur length in our final model, we will compare two models: the preliminary final model without femur length squared vs. the preliminary final model with both the linear femur length variable added as a covariate.

It turns out that the latter model (the one with linear femur length variable added) has a superior AIC to the former model (the one with neither femur length variable included). Thus, we choose the model with both the linear and quadratic femur length variables to be our final model. The VIF scores are lower than that of the initial model, which indicates that multicollinearity has been reduced. The output for these results are shown in Figure 13.

The output for the final model is displayed in Figure 14. The model is:

$$\log\left(\frac{\pi}{1-\pi}\right) = 0.03438 - 2.64965\,\mathrm{TL} + 2.23485\,\mathrm{HL} - 1.44287\,\mathrm{WT} + 1.00213\,\mathrm{KL} + 0.72168\,\mathrm{BH} - 0.02989\,\mathrm{FL} + 0.85894\,\mathrm{FL}^2$$

Note that since we standardized the original variables, an increase in one unit is equal to an increase in one standard deviation. We can then interpret these coefficients as follows:

- For each increase in one standard deviation of total length, the odds ratio of survival decreases by 92.9%
- For each increase in one standard deviation of humerus length, the odds ratio of survival increases by 834.5%
- For each increase in one standard deviation of weight, the odds ratio of survival decreases by 76.4%
- For each increase in one standard deviation of keel of sternum length, the odds ratio of survival increases by 172.4%
- For each increase in one standard deviation of beak and head length, the odds ratio of survival increases by 105.8%

The interpretation is more complicated for the femur length variable since a square term is included. It appears that femur length has a negative effect on survival, but the opposite sign of the quadratic term indicates that the effect becomes less pronounced over time and would even eventually change to being a positive effect, though that occurrence may be well outside the range of the data.

3.2 Model Diagnostics

From examining the diagnostic plots (Figure 16) of the model fit in the final model equation, observations 27 and 40 appear to be potentially unusual, influential or outlying. Observation 27 is an adult sparrow that survived the storm. If we compare the data for this bird (just the variables included in the final model: total length (TL), weight (WT), length of beak and head (BH), length of humerus (HL), length of femur (FL) length of keel of sternum (KL)) to the average for birds who survived the storm, we see that this bird is smaller in every metric (asides weight and total length where it is about average) than the average bird who survived the storm. We would expect bigger and stronger birds to have a better chance of surviving the storm. This bird is unusual because it is exceptionally small compared to its fellow surviving sparrows. The difference between the standardized size measures and the standardized size measures for a surviving bird are reported in Figure 15. The bird is only (slightly) above average in size for two measurements (weight and total length).

Observation 56 is also a potential outlier or influential observation. This sparrow is an adult who did not survive the storm. In the second row of Figure 15, we see that the difference between this birds standardized size measures (just the ones included in the final model) and the standardized size measurements for average sparrow who did not survive are all negative. This bird is very small for a non-surviving sparrow. In some measurements, this bird is 2 to 3 standard deviations below the mean for non surviving sparrows, as seen in Figure 15. This makes it an unusual observation. There is no indication that either of these observations are erroneous and should or can be dropped.

The smoothing line on the residuals vs. fitted plot appears to be varying about zero, which indicates a reasonable model fit.

4 Conclusion

4.1 Overview of the Results

We conclude our study by providing an overview of the results and then discussing potential areas of future study. As seen in the previous section, our final model indicates that the important variables affecting the survival of the sparrows in this incident are total length, humerus length, weight, keel of sternum length, beak and head length, and femur length. Though there is some multicollinearity present among these variables, which are all measures of size, it is better than our initial model in which we included all the variables that were collected by the researcher. Our preliminary final model was chosen using the AIC criterion, which balances model size against explanatory power. Since this model included a square term, we decided to add the corresponding linear term to the model.

Though this set of variables is far from exhaustive, our final model does a reasonably good job of fitting the data. Using the generalized likelihood ratio test, we see that our model does not perform significantly worse than the saturated model that fits a parameter to each observation (which in this case is a perfect fit). Thus, we feel confident in saying that we achieved our goal in finding a model to explain sparrow survival under the conditions that were present in this study.

That being said, we must be careful in applying any of the results of this analysis to broader situations. The conditions that resulted in these sparrows being collected for research were very specific, and the results here may not be indicative of sparrow health in general. Nevertheless, we have identified certain variables as being of particular interest, and further research may use this information as a starting point.

4.2 Further Study

This study may be supplemented by a principal components analysis. Since there is considerable multicollinearity among the original set of variables, perhaps it is possible to find a few linear combinations of the variables that explains most of the variation in the dataset. We could then run a regression on these principal components to determine which ones are significant in explaining survival.

Additional studies could include more variables (for instance, gender) and study the birds over a longer period of time. For example, characteristics on sparrows can be collected at birth and at regular time intervals throughout their lives, and researchers could then study their lifespan and/or survival rate over time. The age variable proved to be insignificant in this study, but perhaps with more detailed data (more levels than just ”juvenile” and ”adult”) this variable could be included and help explain variation in survival.

Appendix

```
          TL        AE        WT        BH        HL        FL        TT        SK        KL
TL 1.0000000 0.6216688 0.5251925 0.2971997 0.3873714 0.4188333 0.3714058 0.4051501 0.3225277
AE 0.6216688 1.0000000 0.4990499 0.3740344 0.7340034 0.6961208 0.6271673 0.3947013 0.4623729
WT 0.5251925 0.4990499 1.0000000 0.4208610 0.4675116 0.4440005 0.4689626 0.3480838 0.4081211
BH 0.2971997 0.3740344 0.4208610 1.0000000 0.5189299 0.5335587 0.5193747 0.4028379 0.4638314
HL 0.3873714 0.7340034 0.4675116 0.5189299 1.0000000 0.8440408 0.7373573 0.4371073 0.4860805
FL 0.4188333 0.6961208 0.4440005 0.5335587 0.8440408 1.0000000 0.7909904 0.4258273 0.4555498
TT 0.3714058 0.6271673 0.4689626 0.5193747 0.7373573 0.7909904 1.0000000 0.3547988 0.4193330
SK 0.4051501 0.3947013 0.3480838 0.4028379 0.4371073 0.4258273 0.3547988 1.0000000 0.2452865
KL 0.3225277 0.4623729 0.4081211 0.4638314 0.4860805 0.4555498 0.4193330 0.2452865 1.0000000
```

Figure 1: Correlation matrix of the continuous variables

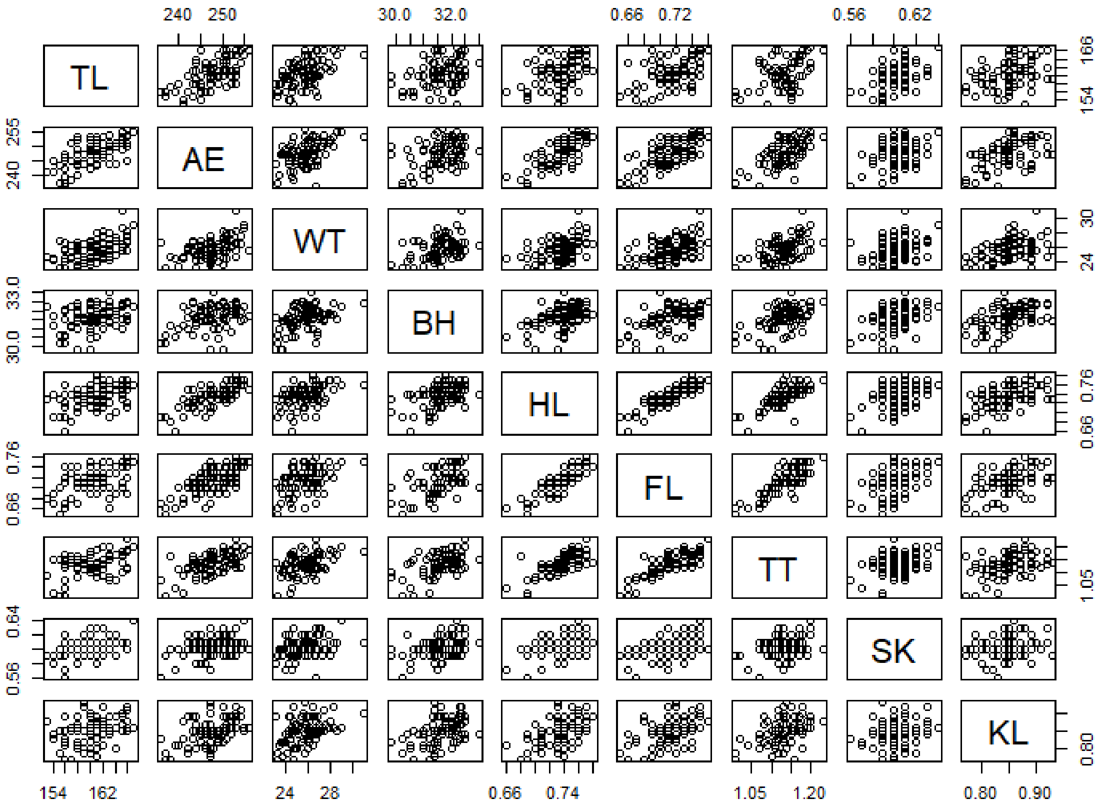


Figure 2: Scatterplot matrix of the continuous variables

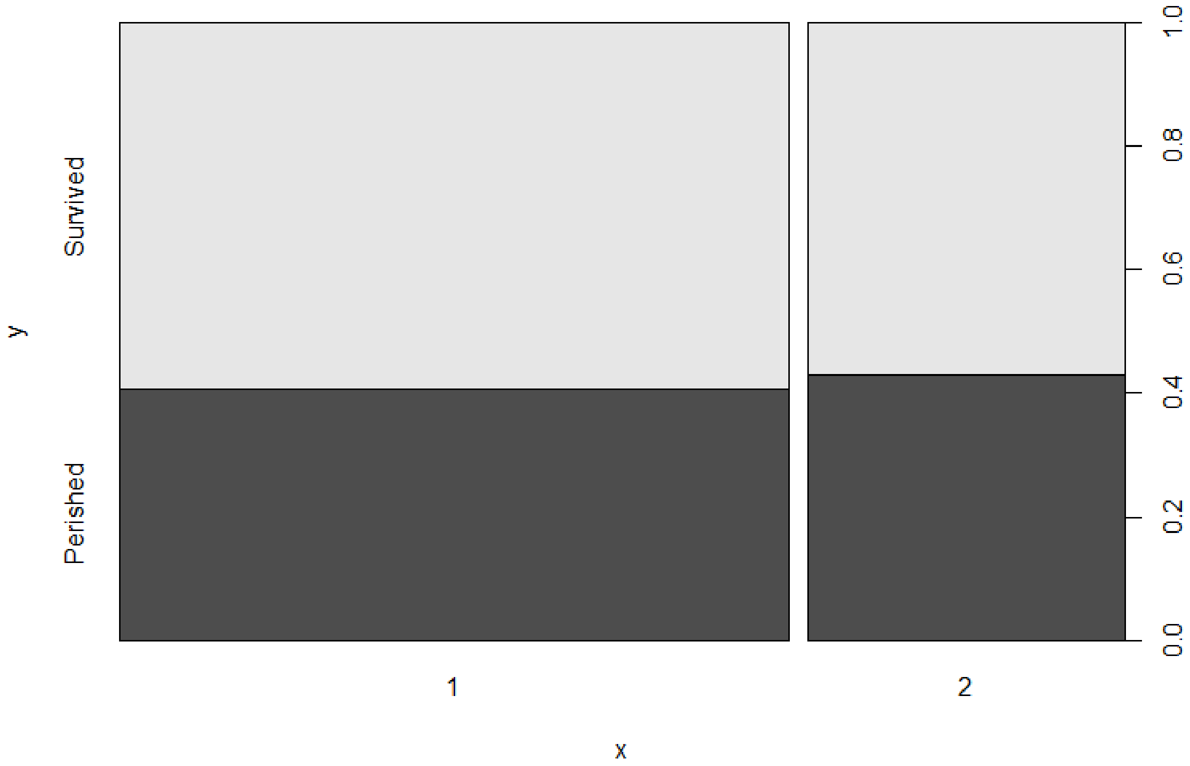


Figure 3: Plot of survival by age group

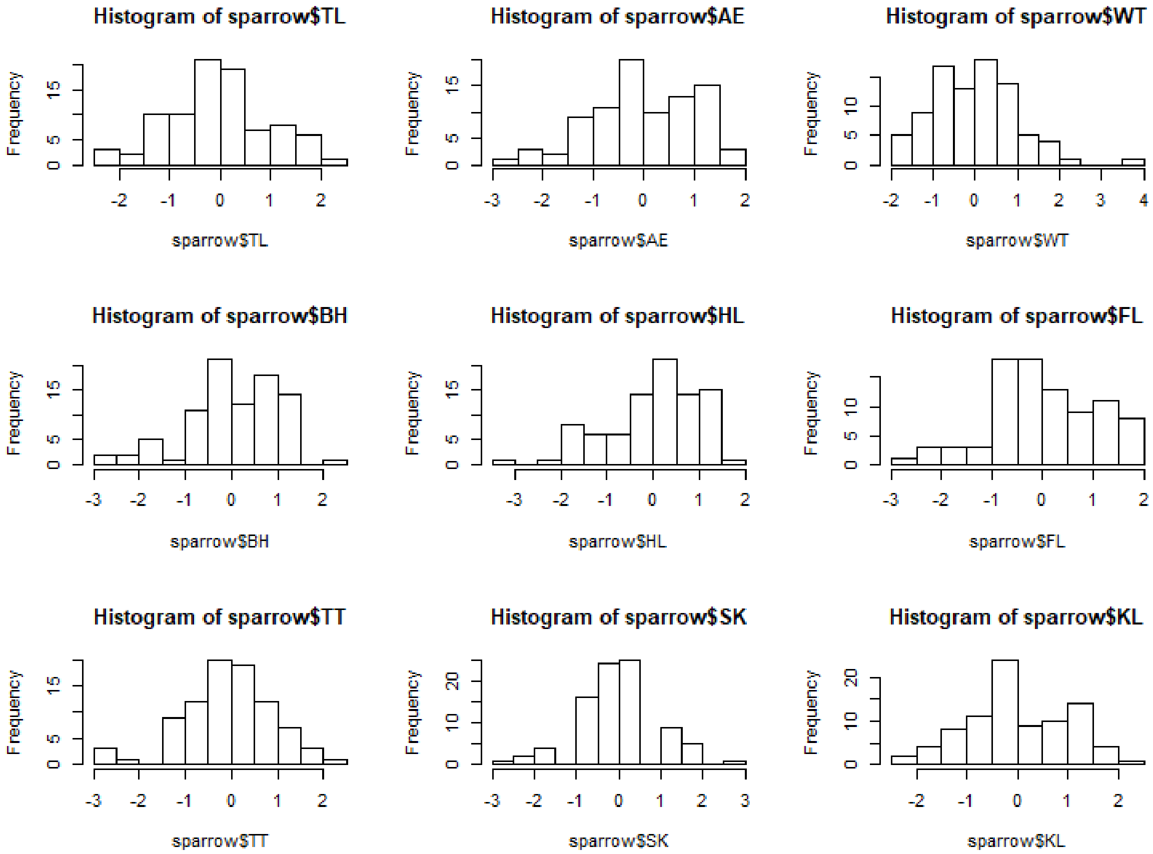


Figure 4: Histograms of the univariate distributions of the continuous variables

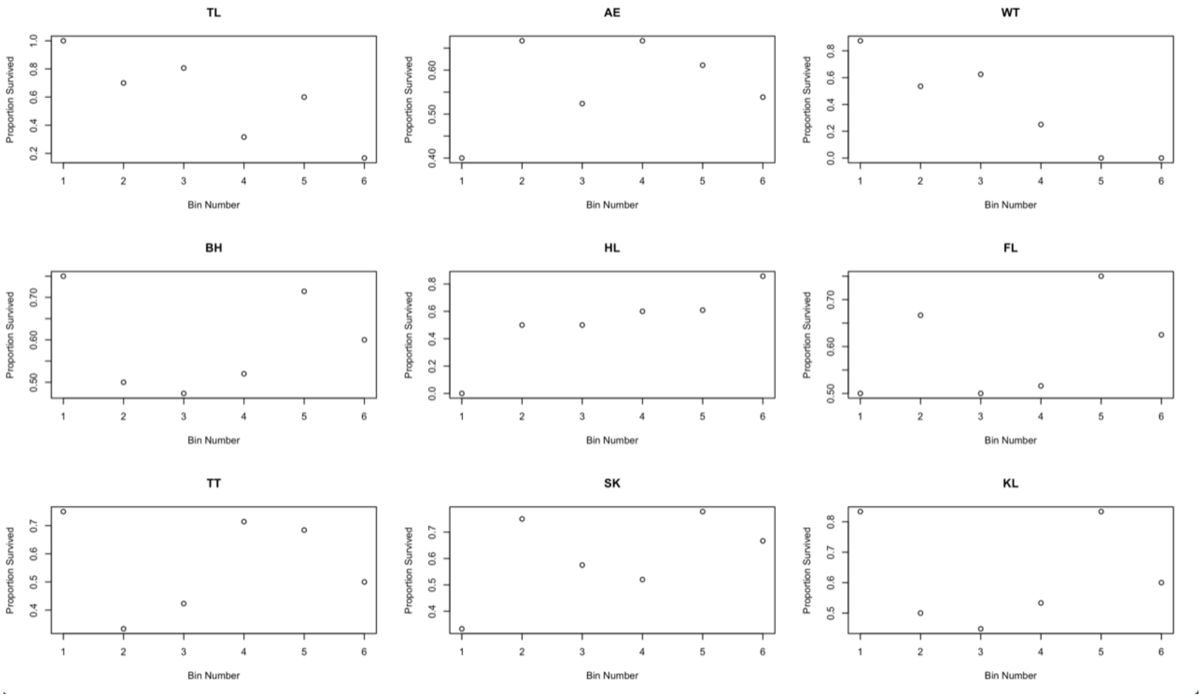


Figure 5: Plots of survival proportion by each predictor

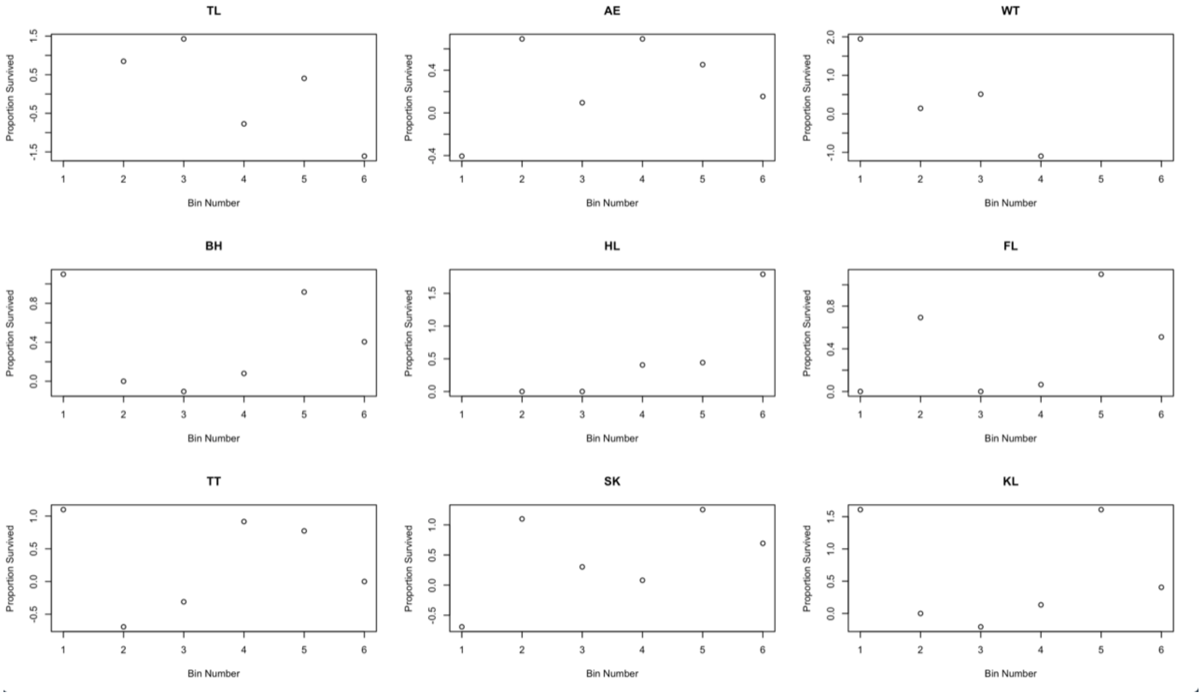


Figure 6: Plots of logit survival proportion by each predictor

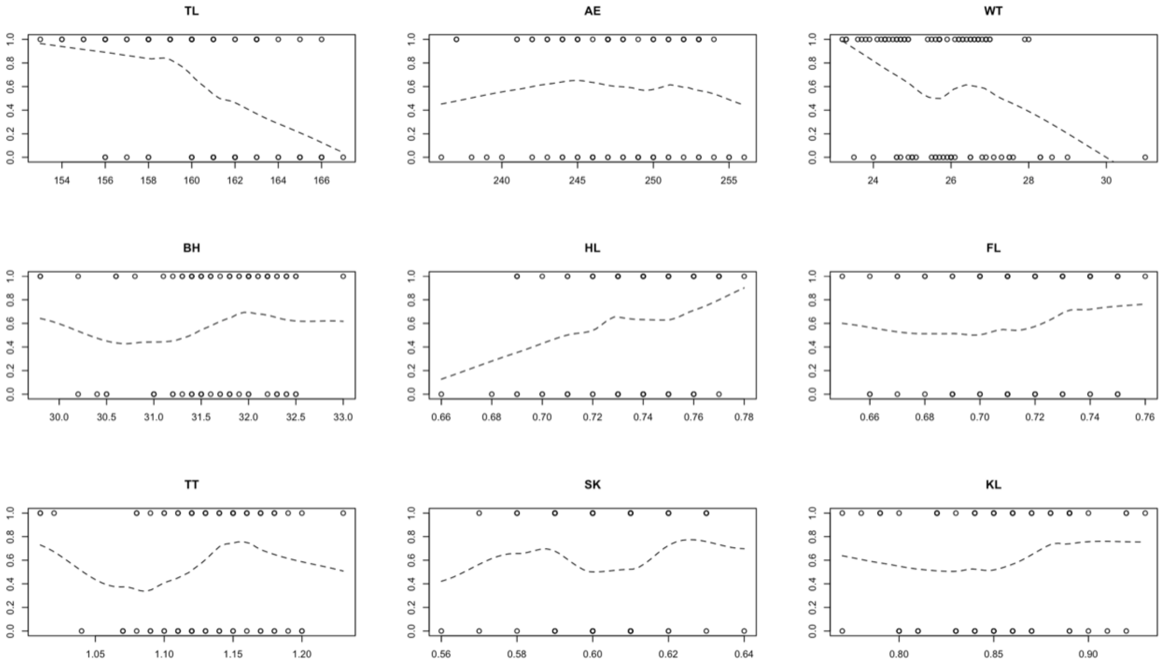


Figure 7: Plots of survival (binary) by each predictor

```
Call:
glm(formula = STATUS ˜ ., family = "binomial", data = sparrow)

Deviance Residuals:
Min       1Q   Median       3Q      Max
-2.2252  -0.5232   0.1397   0.5131   2.0134

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.6510     0.3917   1.662 0.096493 .
AG2           0.1063     0.6825   0.156 0.876225
TL           -2.3598     0.6078  -3.883 0.000103 ***
AE            0.3664     0.5588   0.656 0.512060
WT           -1.2641     0.4863  -2.600 0.009333 **
BH            0.3733     0.3825   0.976 0.329131
HL            1.3039     0.7227   1.804 0.071176 .
FL           -0.1602     0.7649  -0.209 0.834096
TT            0.2075     0.5772   0.360 0.719210
SK            0.3106     0.3936   0.789 0.430037
KL            0.8422     0.4303   1.957 0.050326 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 118.01  on 86  degrees of freedom
Residual deviance:  65.92  on 76  degrees of freedom
AIC: 87.92

Number of Fisher Scoring iterations: 6

> vif(initmodel)
      AG       TL       AE       WT       BH       HL       FL       TT       SK       KL
1.070816 2.714908 3.274337 2.221337 1.892614 4.971644 6.064703 4.024184 1.280141 1.889075
```

Figure 8: R output for the initial model

```
Analysis of Deviance Table

Model 1: STATUS ~ 1
Model 2: STATUS ~ AG + TL + AE + WT + BH + HL + FL + TT + SK + KL
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        86     118.01
2        76      65.92 10   52.088 1.099e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Analysis of Deviance Table

Model 1: STATUS ~ AG + TL + AE + WT + BH + HL + FL + TT + SK + KL
Model 2: STATUS ~ factor(1:length(sparrow$STATUS))
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        76      65.92
2         0       0.00 76    65.92   0.7887
```

Figure 9: Analysis of Deviance Tables for the Initial Model

```
Call:
glm(formula = STATUS ~ . + I(TL^2) + I(AE^2) + I(WT^2) + I(BH^2) +
I(HL^2) + I(FL^2) + I(TT^2) + I(SK^2) + I(KL^2), family = "binomial",
data = sparrow)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.68231  -0.38721   0.04073   0.41931   2.35259

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.1818947  0.6817276  -0.267 0.789612
AG2          0.0004238  0.9615564   0.000 0.999648
TL          -3.5720406  1.0157238  -3.517 0.000437 ***
AE           0.5362849  0.6935474   0.773 0.439375
WT          -1.1044957  0.6474506  -1.706 0.088024 .
BH           0.8955330  0.5151299   1.738 0.082130 .
HL           1.1487515  1.0111177   1.136 0.255906
FL           0.2506067  0.9768099   0.257 0.797521
TT           0.5668612  0.7807699   0.726 0.467821
SK           0.1733182  0.4546662   0.381 0.703056
KL           1.3746269  0.6660167   2.064 0.039022 *
I(TL^2)     -0.4576694  0.5905145  -0.775 0.438319
I(AE^2)     -0.2297263  0.4630081  -0.496 0.619781
I(WT^2)     -0.1492386  0.2466275  -0.605 0.545101
I(BH^2)      0.2859906  0.2493088   1.147 0.251326
I(HL^2)     -0.6648965  0.5231326  -1.271 0.203732
I(FL^2)      1.8472679  0.7870432   2.347 0.018920 *
I(TT^2)     -0.0587736  0.4953948  -0.119 0.905561
I(SK^2)     -0.0211713  0.3030921  -0.070 0.944312
I(KL^2)      0.6636452  0.4474551   1.483 0.138033
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 118.008  on 86  degrees of freedom
Residual deviance:  51.463  on 67  degrees of freedom
AIC: 91.463

Number of Fisher Scoring iterations: 7

Analysis of Deviance Table

Model 1: STATUS ~ AG + TL + AE + WT + BH + HL + FL + TT + SK + KL
Model 2: STATUS ~ AG + TL + AE + WT + BH + HL + FL + TT + SK + KL + I(TL^2) +
I(AE^2) + I(WT^2) + I(BH^2) + I(HL^2) + I(FL^2) + I(TT^2) +
I(SK^2) + I(KL^2)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        76     65.920
2        67     51.463  9   14.458   0.1069
```

Figure 10: R output for quadratic model

```
> vif(fitquad)
      AG       TL       AE       WT       BH       HL       FL       TT       SK       KL  I(TL^2)
1.614328 4.995932 3.188361 3.332696 2.422912 6.241617 5.637711 3.604584 1.527559 2.808544 2.316878
 I(AE^2)  I(WT^2)  I(BH^2)  I(HL^2)  I(FL^2)  I(TT^2)  I(SK^2)  I(KL^2)
2.603950 1.686691 1.544045 4.380960 5.876402 2.601687 1.593285 2.129911
```

Figure 11: VIF scores for the quadratic model

```
Step:  AIC=73.39
STATUS ~ TL + HL + WT + KL + I(FL^2) + BH

          Df Deviance     AIC
<none>         59.395  73.395
+ I(HL^2)  1   57.454  73.454
+ TL:HL    1   57.766  73.766
+ I(AE^2)  1   58.297  74.297
+ I(KL^2)  1   58.426  74.426
+ TL:KL    1   58.489  74.489
+ I(BH^2)  1   58.512  74.512
+ AE       1   58.531  74.531
+ SK       1   58.609  74.609
+ BH:HL    1   58.734  74.734
+ I(SK^2)  1   58.866  74.866
+ TT       1   58.942  74.942
- BH       1   62.982  74.982
+ WT:BH    1   58.995  74.995
+ AG       1   59.086  75.086
+ I(TL^2)  1   59.094  75.094
+ TL:WT    1   59.102  75.102
+ HL:KL    1   59.168  75.168
+ I(TT^2)  1   59.222  75.222
+ WT:KL    1   59.292  75.292
+ BH:KL    1   59.347  75.347
+ I(WT^2)  1   59.376  75.376
+ WT:HL    1   59.376  75.376
+ FL       1   59.393  75.393
+ TL:BH    1   59.394  75.394
- KL       1   65.394  77.394
- I(FL^2)  1   67.214  79.214
- WT       1   71.461  83.461
- HL       1   80.747  92.747
- TL       1   90.405 102.405

Call:  glm(formula = STATUS ~ TL + HL + WT + KL + I(FL^2) + BH, family = "binomial",
data = sparrow)

Coefficients:
(Intercept)           TL            HL            WT            KL       I(FL^2)            BH
0.03504     -2.65277       2.21383     -1.44229       1.00078       0.85832       0.71898

Degrees of Freedom: 86 Total (i.e. Null);  80 Residual
Null Deviance:     118
Residual Deviance: 59.39  AIC: 73.39
```

Figure 12: Result of the stepwise procedure

```
> noFL=glm(formula = STATUS ~ TL + HL + WT + KL + BH, family = "binomial",
+          data = sparrow)
> AIC(noFL)
[1] 79.21435
> bothFL=glm(formula = STATUS ~ TL + HL + WT + KL + BH + FL + I(FL^2), family = "binomial",
+          data = sparrow)
> AIC(bothFL)
[1] 75.39255
> vif(bothFL)
      TL       HL       WT       KL       BH       FL  I(FL^2)
2.755152 5.077029 1.925809 1.686670 1.837898 3.515553 1.839272
```

Figure 13: Comparing the models with and without femur length

```
Call:
glm(formula = STATUS ~ TL + HL + WT + KL + BH + FL + I(FL^2),
family = "binomial", data = sparrow)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.09194  -0.43063   0.09373   0.50091   2.62510

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.03438    0.41931   0.082  0.93466
TL          -2.64965    0.66894  -3.961 7.46e-05 ***
HL           2.23485    0.78513   2.846  0.00442 **
WT          -1.44287    0.48321  -2.986  0.00283 **
KL           1.00213    0.45123   2.221  0.02636 *
BH           0.72168    0.39578   1.823  0.06824 .
FL          -0.02989    0.67265  -0.044  0.96456
I(FL^2)      0.85894    0.36872   2.329  0.01983 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 118.008  on 86  degrees of freedom
Residual deviance:  59.393  on 79  degrees of freedom
AIC: 75.393

Number of Fisher Scoring iterations: 6

Analysis of Deviance Table

Model 1: STATUS ~ TL + HL + WT + KL + BH + FL + I(FL^2)
Model 2: STATUS ~ factor(1:length(sparrow$STATUS))
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        79     59.393
2         0      0.000 79   59.393   0.9513
```

Figure 14: R output for the final model

```
      HL      TL        WT         KL          BH        FL
27: -1.255516  0.232497   0.1571303 -1.826644  -0.1469722 -0.667049
56  -2.996246 -1.898215  -1.169633  -0.9869181 -1.65701   -2.062835
```

Figure 15: Data for the outliers

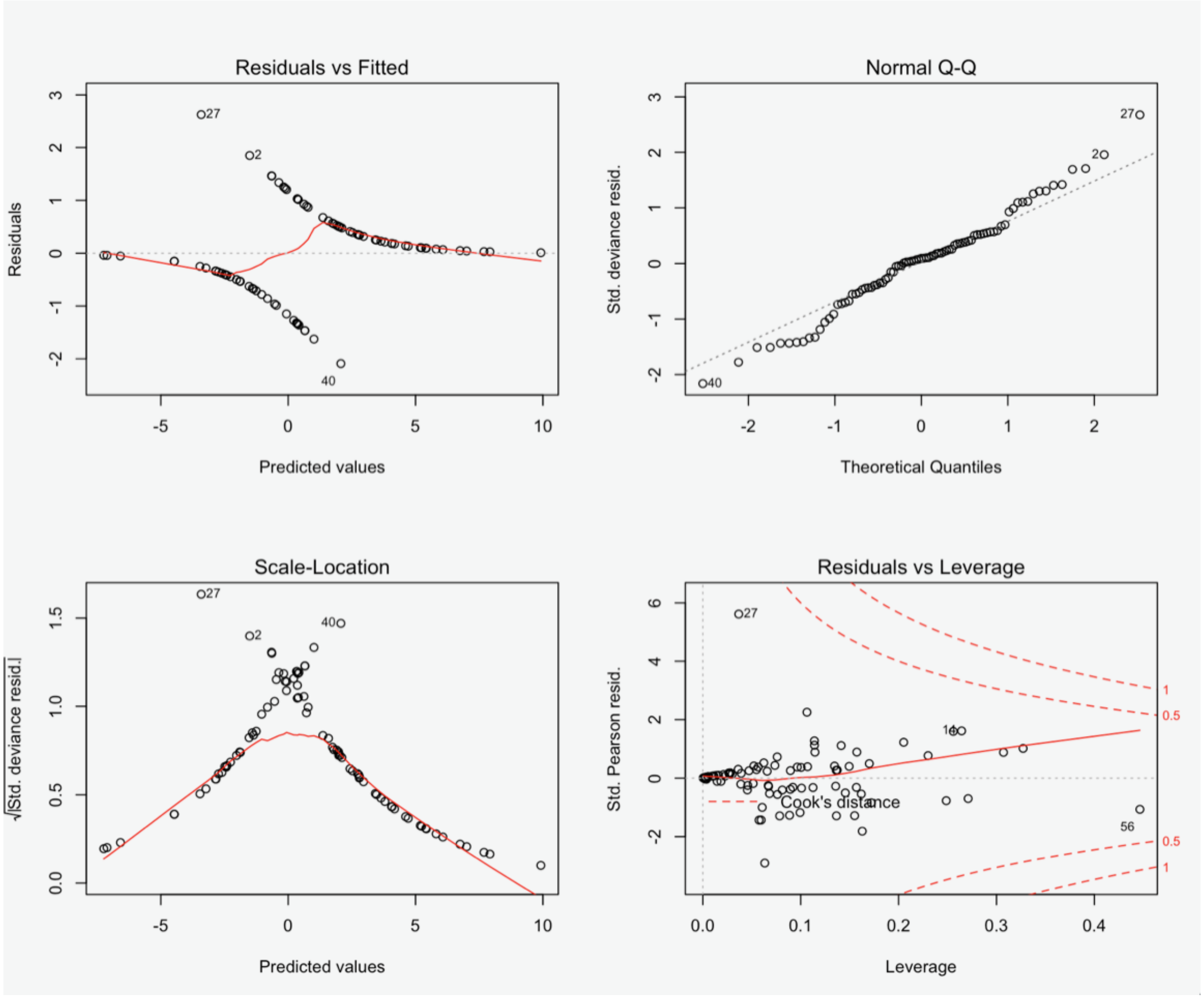


Figure 16: Diagnostic Plots for the Final Model

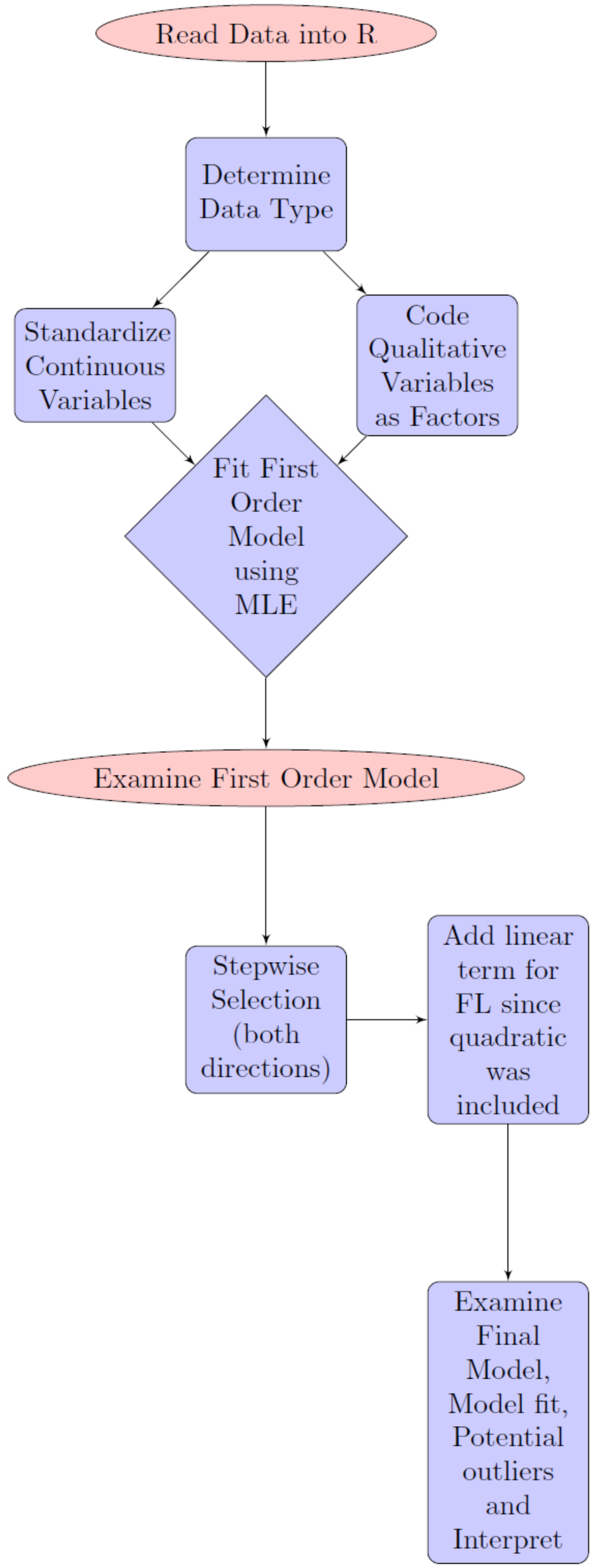
Code

```
k = read.table("/Users/douglasturner/Desktop/survival_sparrow.txt", header=TRUE)
num=rep(0,dim(k)[1])
title =c("STATUS", "AG"  ,"TL", "AE",  "WT"  ,"BH", "HL" , "FL"  , "TT", "SK", "KL")
for (i in c(1:dim(k)[1])) {
  if (k[i,1]=="Survived") {
    num[i]=1
  }
}

k$AG = factor(k$AG)

#Figure 1: Correlation matrix
cor(k[,3:11])
# Figure 2: Scatterplot matrix
plot(k[,3:11])
# Figure 3:
par(mfrow=c(1,1))
plot(k$AG, k$STATUS)

k$STATUS = factor(k$STATUS)
#Figure 5
title =c("STATUS", "AG"  ,"TL", "AE",  "WT"  ,"BH", "HL" , "FL"  , "TT", "SK", "KL")
a = 7
lower = c()
upper=c()
diff=c()
s=matrix(0,a,11)
for (j in c(3:11)) {
  lower[j] = min(k[,j])
  upper[j] = max(k[,j])
  diff[j] = (upper[j]-lower[j])/a
  s[,j] = seq(from=0, to =a-1)*diff[j] +lower[j]
}
#Regular Bin Plot
par(mfrow=c(3,3))
for (j in c(3:11)) {
  g = data.frame(k[,j], bin=cut(k[,j], a-1, include.lowest=TRUE))
  ans = data.frame(num,g[,2])
  ans = table(ans)[2,]/(table(ans)[1,]+table(ans)[2,])
  plot(seq(1:(a-1)),ans, xlab="Bin Number", ylab = "Proportion Survived", main=title[j])
  #points(loess.smooth(seq(1:(a-1)), ans),type='l' ,lty=2)
}

#Figure 6: Logit
#Logit Tranform
par(mfrow=c(3,3))
for (j in c(3:11)) {
  g = data.frame(k[,j], bin=cut(k[,j], a-1, include.lowest=TRUE))
  ans = data.frame(num,g[,2])
  ans = table(ans)[2,]/(table(ans)[1,]+table(ans)[2,])
  plot(seq(1:(a-1)),log((ans)/(1-ans)), xlab="Bin Number", ylab = "Proportion Survived", main=title[j])
  #points(loess.smooth(seq(1:(a-1)), ans),type='l' ,lty=2)
}
```

```
#Figure 7
par(mfrow=c(3,3))
for (i in c(3:11)) {
  plot(k[,i], num, xlab = " ", ylab=" ", main=title[i])
  points(loess.smooth(k[,i], num),type=’l’,lty=2)
}


#Model Fitting:
#Model with first order terms
k[,3:11] = scale(k[,3:11])
# Figure 4: Histogram
par(mfrow=c(3,3))
hist(k$TL)
hist(k$AE)
hist(k$WT)
hist(k$BH)
hist(k$HL)
hist(k$FL)
hist(k$TT)
hist(k$SK)
hist(k$KL)


k=data.frame(k,num)
fit = glm(num~AG+TL+AE+WT+BH+HL+FL+TT+SK+KL, family=binomial(link=logit), data=k)
library("car")

#Figure 8
summary(fit)
vif(fit)

#Null model
k = k[,-1]
fitnull = glm(num~1, family=binomial(link=logit), data=k)
#Saturated Model
a = factor(1:length(k$AG))
fitsat = glm(num~a, family=binomial(link=logit), data=k)
#Figure 9
anova(fitnull, fit, test="Chisq")
anova(fit,fitsat, test="Chisq")

#Quadratic Model
#Figure 10
fitquad = glm(num~+AG+TL+AE+WT+BH+HL+FL+TT+SK+KL+I(TL^2)+I(AE^2)+I(WT^2)+I(BH^2)+I(HL^2)+I(FL^2)+I(TT^2
summary(fitquad)
anova(fit,fitquad, test="Chisq")
vif(fitquad)
```

```
#Model with all second order and interactions

fitall = glm(num~.^2+AG+TL+AE+WT+BH+HL+FL+TT+SK+KL+I(TL^2)+I(AE^2)+I(WT^2)+I(BH^2)+I(HL^2)+I(FL^2)+I(TT
summary(fitall)
anova(fitall, fit, test="Chisq")
library("MASS")
#Figure 11
stepAIC(fitnull,c(lower=formula(fitnull), upper= formula(fitall)), direction=c("both"), k=2)
#Figure 12
fitFL2 = glm(num~TL+HL+WT+KL+FL+I(FL^2)+BH, family=binomial(link=logit),data=k)
fitnoFL =  glm(num~TL+HL+WT+KL+BH, family=binomial(link=logit),data=k)
AIC(fitFL2)
AIC(fitnoFL)
vif(fitFL2)


#Final Model
#Figure 13
finalfit = glm(num~TL+HL+WT+KL+FL+I(FL^2)+BH, family=binomial(link=logit),data=k)
summary(finalfit)
vif(finalfit)
par(mfrow=c(2,2))

dev.off()
anova(finalfit, fitsat, test="Chisq")

#Observations 27 and 40 appear to be influential observations and possible outliers
k[27,c("HL","TL", "WT", "KL", "BH", "FL")]
k[56,c("HL","TL", "WT", "KL", "BH", "FL")]
#Figure 14
k[27,c("HL","TL", "WT", "KL", "BH", "FL")] - colMeans(subset(k[,c("HL","TL", "WT", "KL", "BH", "FL")], 
k[56,c("HL","TL", "WT", "KL", "BH", "FL")] - colMeans(subset(k[,c("HL","TL", "WT", "KL", "BH", "FL")], 
#Figure 15
plot(finalfit)
```

Flowchart


Read Data into R
Determine Data Type
Standardize Continuous Variables
Code Qualitative Variables as Factors
Fit First Order Model using MLE
Examine First Order Model
Stepwise Selection (both directions)
Add linear term for FL since quadratic was included
Examine Final Model, Model fit, Potential outliers and Interpret