Project2: House sparrow survival data.

Group members:

Jiahuan He

Yibing Luo

Yingjie Li

Abstract:

A logistic regression model is built to explore the relationships between the probability of house sparrow’s survival and their physical characteristics. Step wise selection is applied to find the “best” model. The final model shows that the smaller and the stronger the sparrows, the more likely that they will survive during the winter storm.

**Introduction**

In this project, a logistic regression model is built to explore the relationships between the probability of house sparrow’s survival and their physical characteristics. We hope that the logistic regression model can provide a good predictionof the results of survival probability. In addition, the final model should also have easily interpretable terms, which allows us to infer the relationship between response variables and those features.

**Material**

First of all, we may take a look at the summary of the data. There are 87 observations in the data set. 51 of them are survival, which takes 58.6% of the whole sample. There are ten physical characteristics (potential predictors), one of them is categorical: Age (adult=1, juvenile=2). The other 9 are quantitative: Total length (TL), Alar extent (AE), Weight (WT), Length of beak and head (BH), Length of humerus (HL), Length of femur (FL), Length of tibio-tarsus (TT), Width of skull (SK), Length of keel of sternum (KL). In addition, value 0 = Perished, 1 = survived of Survival.

**Method**

In this project, the multiple logistic regression analysis is built and stepwise selection process is applied to find the “best” model. After the final model is derived, some necessary diagnostics are conducted, e.g. test for the residuals and the goodness-of-link test. Since normal assumption for residuals is not important for logistic regression, variable transformation is not considered in model building.

1. Check the Multicollinearity.

Since most of the predictor variables measure the magnitude of a house sparrow, multicollinearity might exist among those the variables. Therefore, the multicollinearity is checked using Variance Inflation Factor(VIF) before the stepwise selection is applied.

From Figure 1 we find that some of the predictors are highly correlated to each other. However, In terms of VIF, the multicollinearity is not severe among those predictor variables:



In fact, the VIF of all ten predictors are not very large (the maximum VIF is 4.67), which means the multicollinearity is not a problem during this model building.

1. Fit the full model

We fit the full model with all variables and do some initial analysis. The reason why only first order terms are considered as the full model is stated in part 3).

From Figure 2 we find that there are only a few variables with small P-values. It indicates that some variables in the full model are redundant. Thus, a model selection process is needed to find the “best” model.

1. Model Selection

First of all, we need to decide which model is our full model. Since there are only 87 observations in the data set, which is relatively small comparing to the number of candidate predictor variables we have. And if the second order terms or the interaction terms are about to be included, then the observation-variable ratio would be even smaller, which violate the rule of thumb that a model should have at least 6 to 10 observations per variable. Therefore, we select several models to be the full models to do stepwise selection.

1. Model only involves first-order terms.
2. Model involves first-order terms and all second interaction terms.
3. Model involves first-order terms and second-order terms, the second order terms based on result of a.

AIC criterion is used to do stepwise selection.

For a, the variables TL, WT, HL and KL are selected.

For b, the result is the same as a.

For c, the second order of TL, WT, HL and KL are added to the full model and the final model that we end up with is also the same

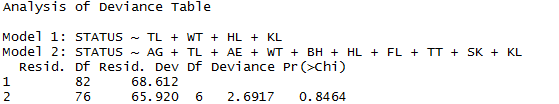
Thus, the final model function is as follow,

And the summary of the model is as Figure 8.

Since all three different full model finally yields the same best model, full model a. is chosen as the baseline for goodness of fit test.

1. Model diagnose
2. Test the goodness of fit of the final model

We will test the final model’s goodness of fit against the full model:



P-value is 0.84, which means the reduced model fitted the data almost as good as the full model. Thus, it is reasonable to exclude the other predictor variables.

1. Residual plot:

From Figure 4, we can see the lowess line is around Y=0, which means the model is meaningful.

1. Detection of outlier and influential point

First the Pearson residuals are standardized. And we draw the plot of the Pearson residuals. Empirically, those cases with the absolute value of Pearson residuals greater than 2 are considered to be outlier in Y. From Figure5, cases 70 and 27 are identified. We also draw half-normal plot in Figure 6, and the conclusion above is confirmed.

We use leverage to find outlier in X. Plot is shown as Figure 7. And case 81 is considered as outlier in X.

Furthermore, whether these outliers are influential points is checked. We take those three outliers out one by one, and calculate the average changing percent of all coefficients. And the result is as follow:



It turns out that case 27 and 76 are likely to be the influential points, thus we delete them from the data and refit the model. The new model is:

1. Leave-one-out cross-validation

To justify our deletion of those two observations, leave-one-out cross-validation is applied and mean squared error(MSE) is calculated. The MSE before those two observations are excluded is 0.15, and 0.1327 after the deletion. This indicates that those two observations seems to unduly affect our model and it is necessary to delete them from the training data.

**Conclusion and Discussion**

Based on the logistic regression model we fit. Four of the ten predictor variables are included in the model without any interaction terms or high-order term, which makes the model easily interpretable. **The signs of the estimated coefficients of “WT” and “TL” are all negative. Since WT measures the weight of house sparrow while TL shows the total length. This indicates that the smaller the house sparrow, the more likely that it will survive in a severe winter storm. Biologically speaking, this is reasonable. Sparrow with a smaller body will consume less food and find a shelter more easily. The coefficients of HL(Length of humerus) and KL(Length of keel of sternum) are positive. This is also understandable, since those two variables measure the robustness of a sparrow, and obviously, a stronger sparrow is more likely to survive. Generally, our final model is interpretable.**

However, the size of our training data is small. Thus, we will be more comfortable to draw the above conclusion with more observations.

**Appendix:**

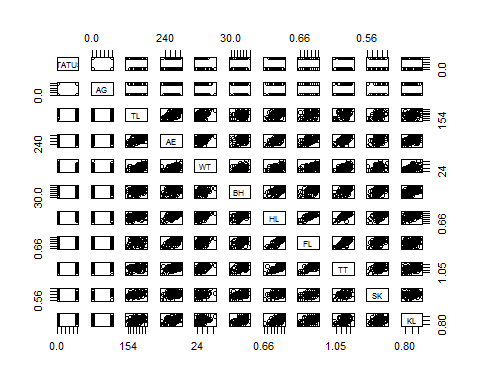


Figure 1 Scatter Plot

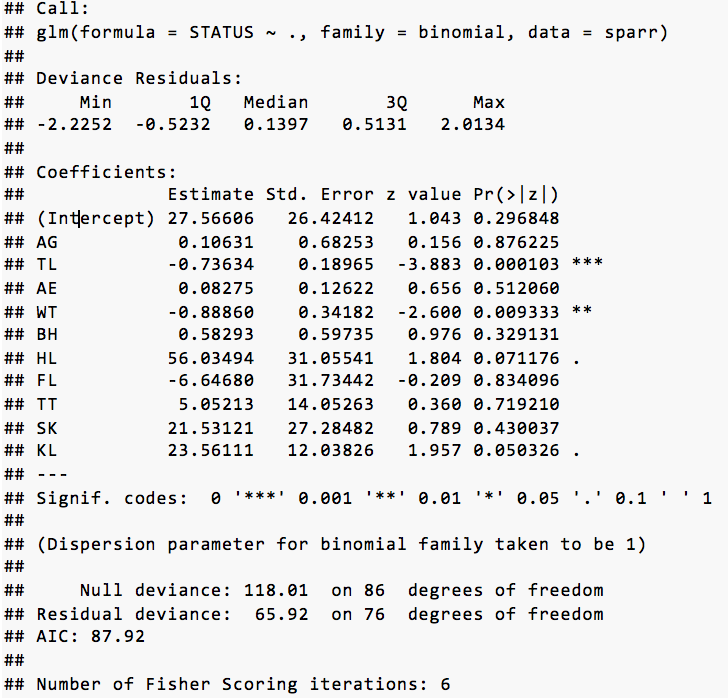


Figure 2 Summary of the Full Model

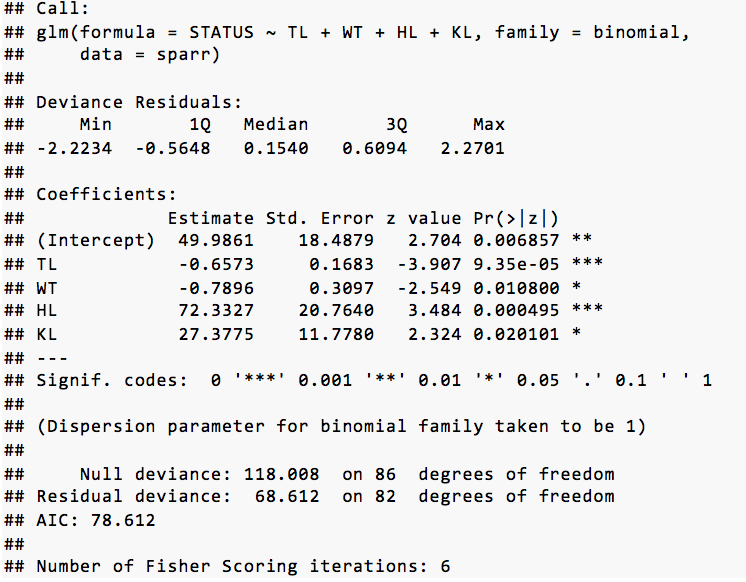


Figure 3 Summary of Stepwise Selection

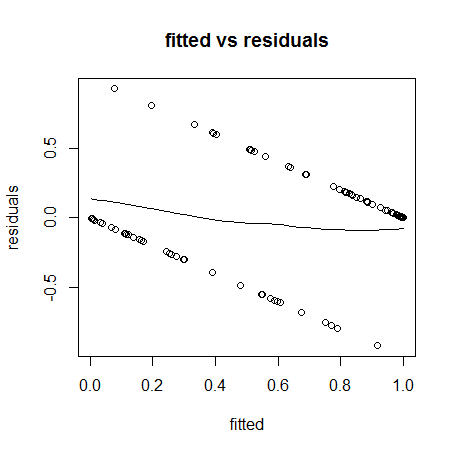


Figure 4

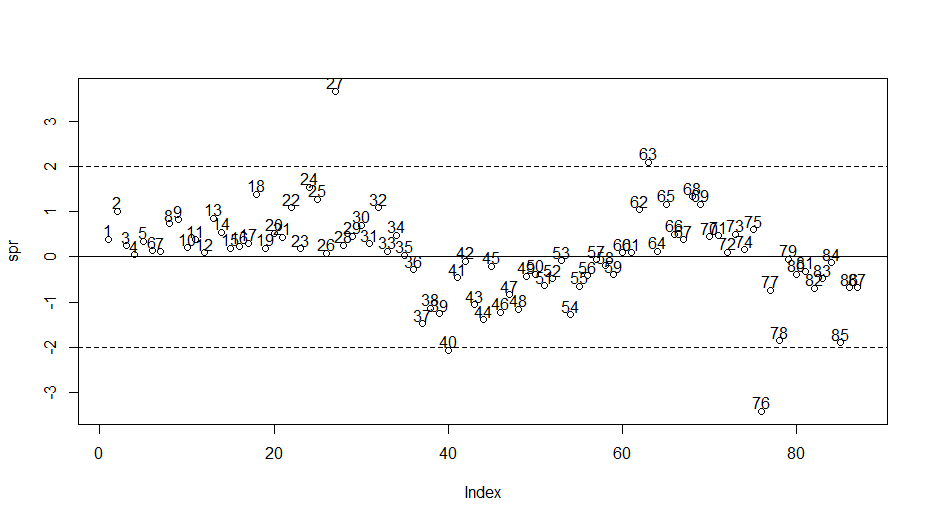


Figure 5

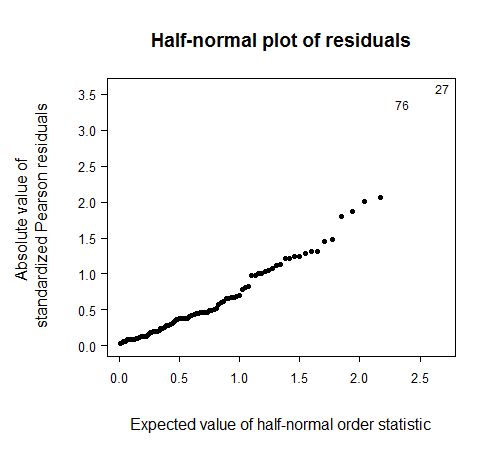


Figure 6

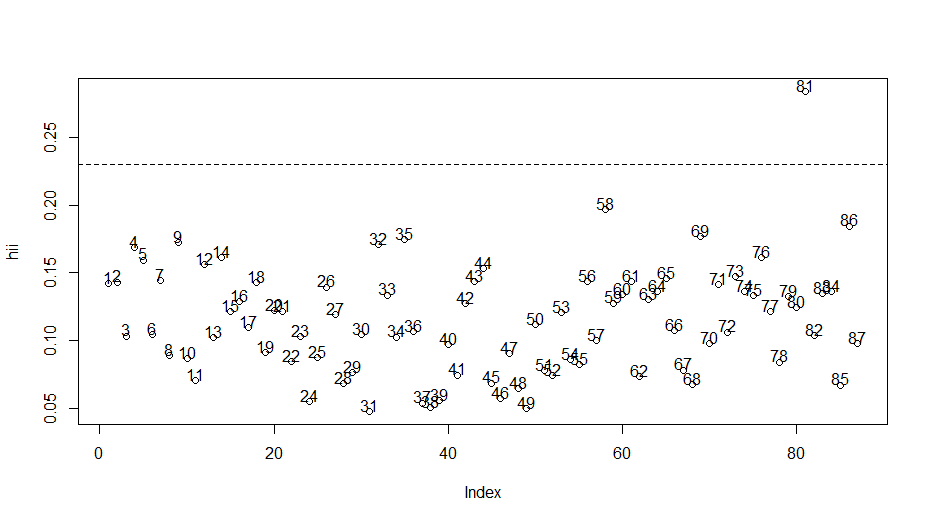


Figure 7

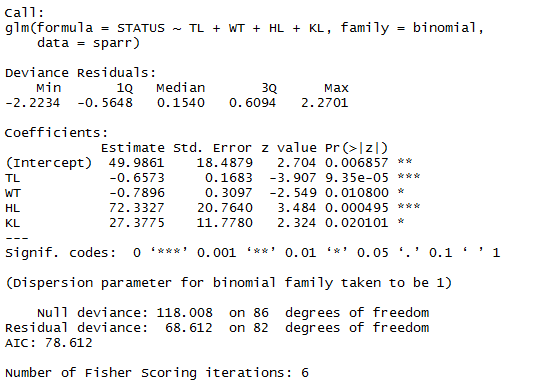


Figure 8