Week 6 Lab Report

Nguyen Tien Anh Quach

2024-03-07

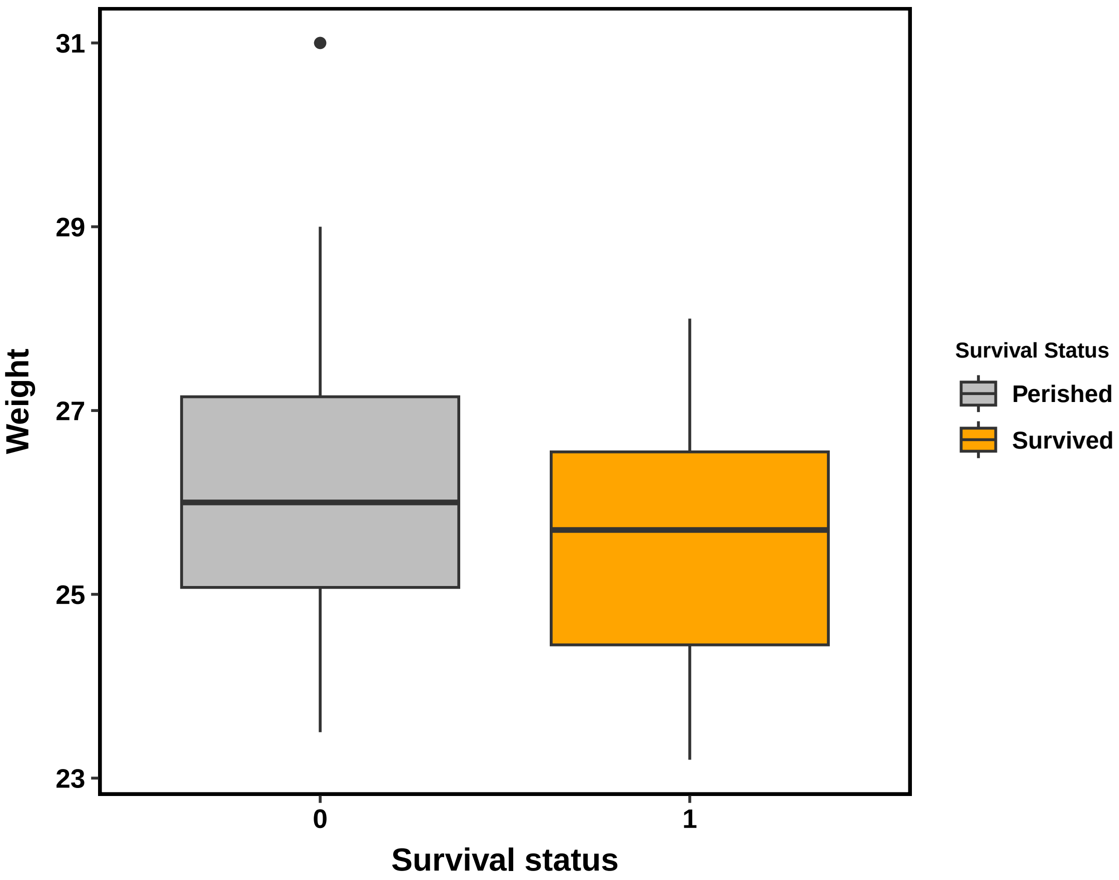
Question 1

1. The simple logistic regression model was:

$$\text{stat} \sim \text{WT}$$

For stat, a value of 0 means that a bird is perished and a value of 1 means that a bird survived.

2. Boxplots of survival across weights



Birds' weight is significantly and negatively associated with the survival odds. For every one unit increase in weight, the log odds of survival decreases by 0.42. In other words, as the odds ratio is 0.65, for every increase of 1 unit in weight, the chance for a bird to survive goes down by 35%.

3. Wald confidence interval for β_1

Calculated by hands: (-0.765, -0.084)

```
## [1] -0.765048
```

```
## [1] -0.083752
```

Calculated using confint: (-0.788, -0.101)

```
##                  2.5 %     97.5 %
## (Intercept)  2.9370104 20.7606136
## WT          -0.7883819 -0.1007059
```

4. Likelihood ratio test:

From the result, weight has a likelihood ratio chi-squared statistic of 6.7595 (df = 1, p = 0.009). Therefore, the result suggests that weight is statistically significant in predicting the survival outcome.

5. Check model assumptions

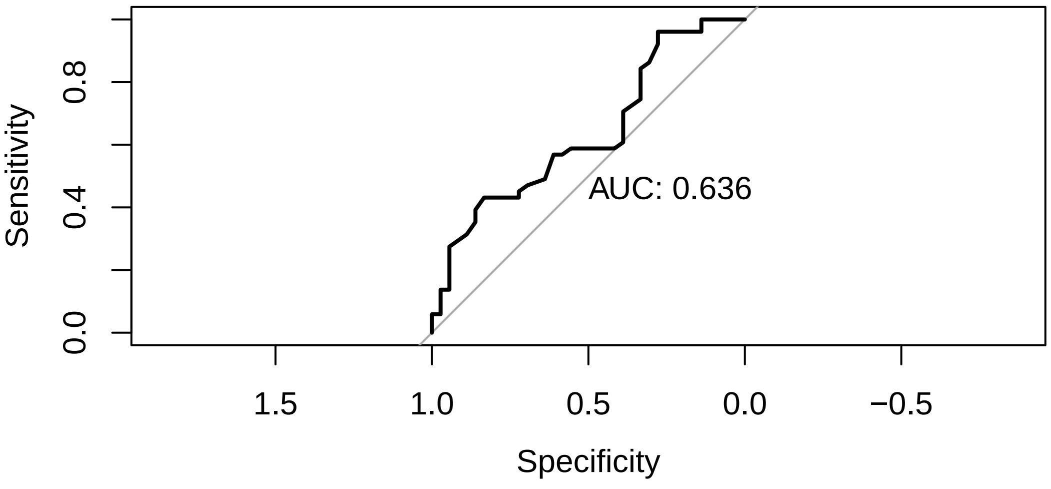
a. Linearity:

The Box-Tidwell test yielded a non-significant p-value of 0.76. Therefore, we fail to reject the null hypothesis and the linearity assumption is met!

b. Influential observation:

There is no influential observation as no observation has the standardized residuals more than 3 or less than -3.

c. Model performance:



This simple model between survival status and weight is a relatively good model, as the AUC = 0.636.

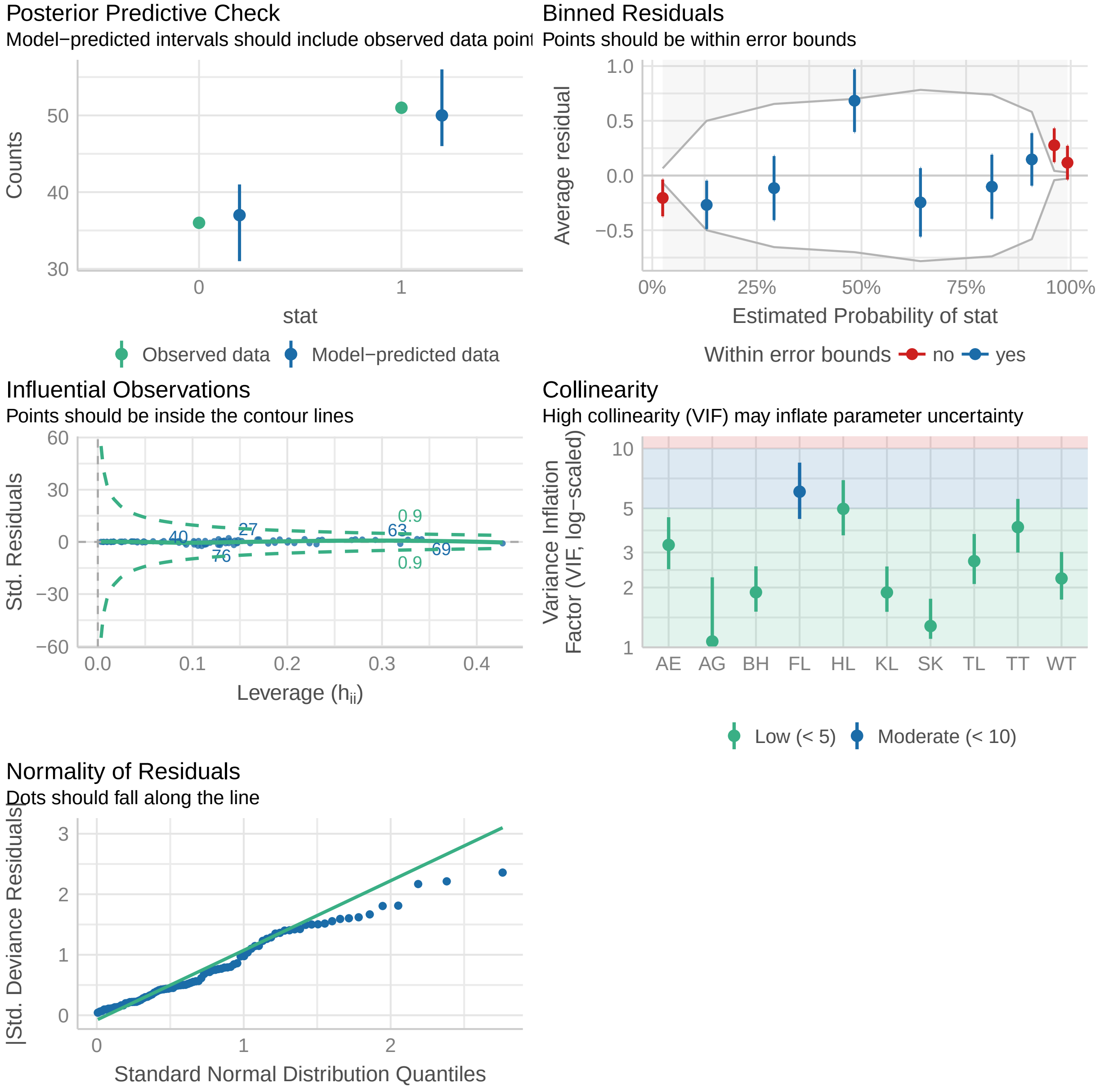
Question 2

1. The full logistic regression model was:

$$\text{stat} \sim \text{WT} + \text{AG} + \text{TL} + \text{AE} + \text{BH} + \text{HL} + \text{FL} + \text{TT} + \text{SK} + \text{KL}$$

The model resulted in only significant effects of weight ($p = 0.009$) and total length ($p < 0.001$) on survival odds. Having 10 variables in the logistic regression model may result in multicollinearity, thus, potentially mask the potentially effects of other variables on birds' survival odds.

Checking for multicollinearity



Results from VIF and visual check of model assumptions revealed that most variables have their VIFs less than 5, except for FL (femur length) with VIF of 6.06. Therefore, I am going to remove FL from the model and re-run.

Revised full model

$$\text{stat} \sim \text{WT} + \text{AG} + \text{TL} + \text{AE} + \text{BH} + \text{HL} + \text{TT} + \text{SK} + \text{KL}$$

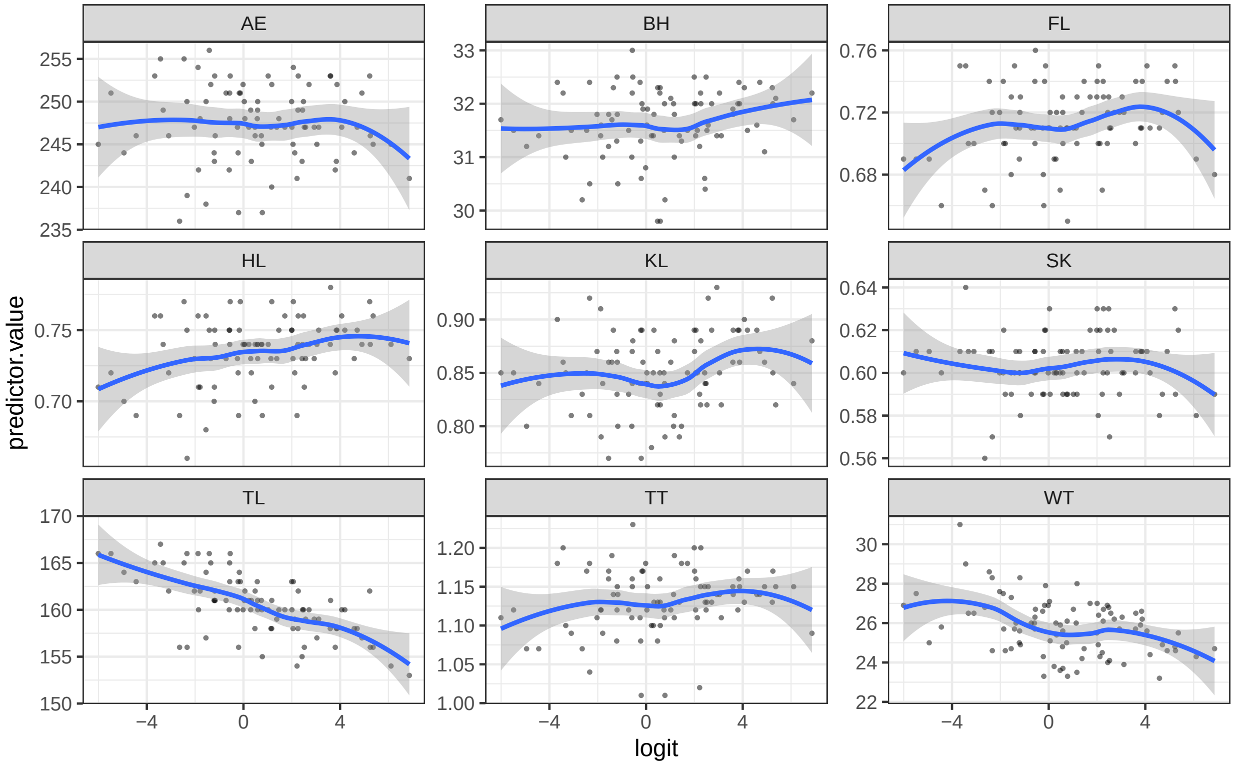
The model results showed that birds' weight, total length, and humerus length are significantly associated with the survival odds. Holding all other variables constant:

For every one unit increase in weight, the log odds of survival decreases by 0.88. In other words, as the odds ratio is 0.41, for every increase of 1 unit in weight, the chance for a bird to survive goes down by 59%.

For every one unit increase in total length, the log odds of survival decreases by 0.74. In other words, as the odds ratio is 0.47, for every increase of 1 unit in weight, the chance for a bird to survive goes down by 53%.

For every one unit increase in humerus length, the log odds of survival increases by 52.7. In other words, as the odds ratio is 8.2e22, for every increase of 1 unit in humerus length, the chance for a bird to survive goes up by many many times (or the probability of surviving is 100%).

2. Checking for linearity



From the visual check, it looks like the relationships between logit and (1) AE, (2) BH, (3) FL, (4) KL, (5) SK, and (6) TT are not linear. Therefore, I am excluding these variables from the model. In addition, I am excluding AG because whenever AG is included in the model, the Box Tidwell test fails. The model now becomes:

$$\text{stat} \sim \text{WT} + \text{TL} + \text{HL}$$

Results from the Box-Tidwell yielded non-significant values for all three predictor variables, WT ($p = 0.93$), TL ($p = 0.84$), and HL ($p = 0.76$).

Re-run the new revised model one more time

$$\text{stat} \sim \text{WT} + \text{TL} + \text{HL}$$

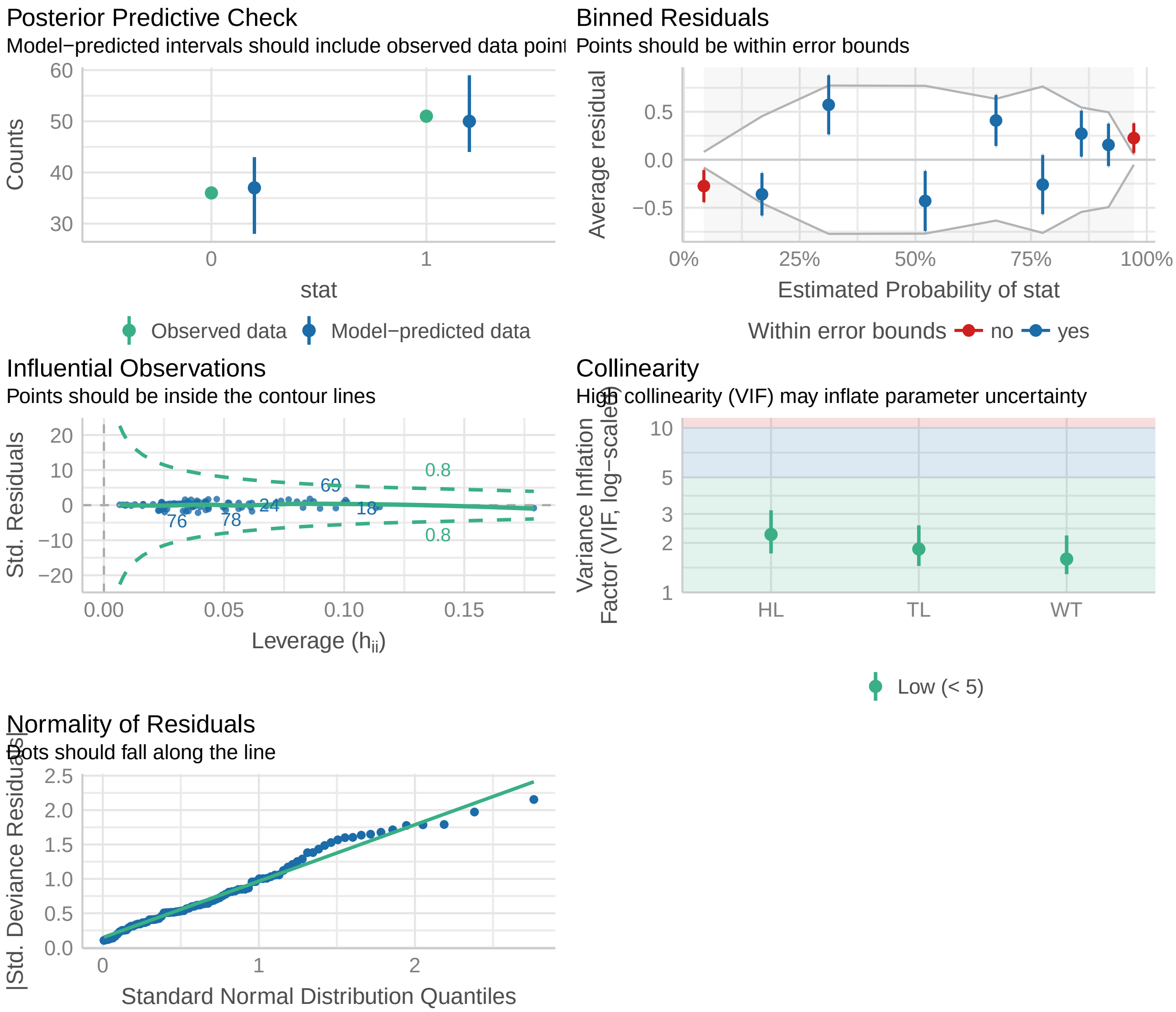
The model results showed that birds' weight, total length, and humerus length are significantly associated with the survival odds. Holding all other variables constant:

For every one unit increase in weight, the log odds of survival decreases by 0.57. In other words, as the odds ratio is 0.57, for every increase of 1 unit in weight, the chance for a bird to survive goes down by 43%.

For every one unit increase in total length, the log odds of survival decreases by 0.54. In other words, as the odds ratio is 0.58, for every increase of 1 unit in total length, the chance for a bird to survive goes down by 42%.

For every one unit increase in humerus length, the log odds of survival increases by 75.5. In other words, as the odds ratio is 5.9e32, for every increase of 1 unit in humerus length, the chance for a bird to survive goes up by many many times (or the probability of surviving is 100%).

Checking other assumptions



Visually checking, it looks like the multicollinearity and normality assumptions are met. In addition, there seems to be no influential observation. Model assumptions are not violated!

3. Cross validation for best model

Using cross validation (codes found in CV for best model R script), the best model was:

$$\text{stat} \sim \text{TL} + \text{HL} + \text{KL} + \text{WT}$$

This was based on several parameters. First, the model has high sensitivity (0.782) and specificity (0.683), although some other models are a bit higher in either parameter. Second, most importantly, the model above has the highest AUC (0.869), which implies that the predictors in that model performed better than the others in other sets of models.

Best model interpretation

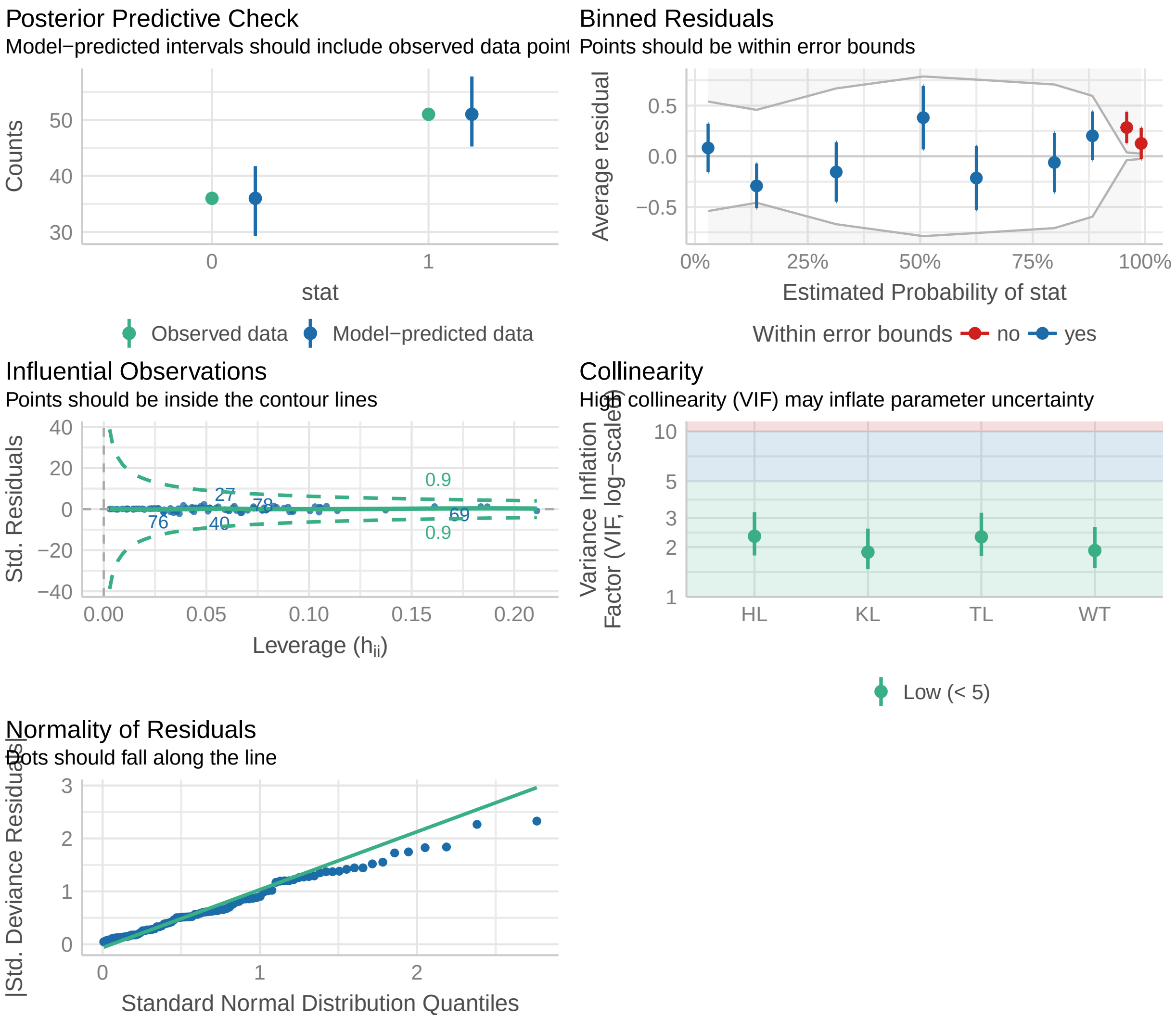
The model results showed that birds' weight, total length, humerus length, and keel of sternum length are significantly associated with the survival odds. Holding all other variables constant:

For every one unit increase in weight, the log odds of survival decreases by 0.79. In other words, as the odds ratio is 0.45, for every increase of 1 unit in weight, the chance for a bird to survive goes down by 55%.

For every one unit increase in total length, the log odds of survival decreases by 0.66. In other words, as the odds ratio is 0.52, for every increase of 1 unit in total length, the chance for a bird to survive goes down by 48%.

For every one unit increase in humerus length and keel of sternum length, the log odds of survival increase by 72.3 and 27.4. In other words, as the odds ratios are 2.59e31 and 7.76e11, for every increase of 1 unit in humerus length and keel of sternum length, the chance for a bird to survive goes up by many many times (or the probability of surviving is 100%).

Checking assumptions



Visually checking, it looks like the multicollinearity and normality assumptions are met. In addition, there seems to be no influential observation. The Box-Tidwell test yielded non-significant results for all variables, WT (p = 0.93), TL (p = 0.99), HL (p = 0.80), and KL (p = 0.36). Model assumptions are not violated!

4. Fit model

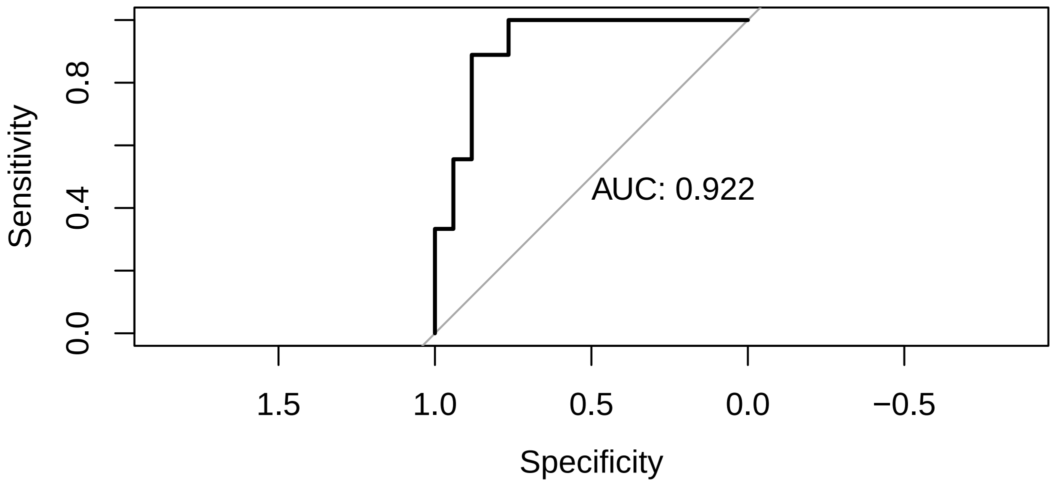
Fit model through 70% of data

$$\text{stat} \sim \text{WT} + \text{TL} + \text{HL} + \text{KL}$$

Assess specificity and sensitivity on 30% data prediction

From the confusion matrix, the accuracy of the model at the cutoff of 0.6 is 0.88, which means that the model makes the accurate prediction of survival outcome 88% of the time. The sensitivity of the model is 0.77, which means that the model correctly predicts a survived bird “survived” 77% of the time. The specificity of the model is 1.0, which means that 0% of the time, the model predicts that a bird survived, when it actually did not.

Plot ROC curve on test data



This model is a great model. Even when it was trained on 70% of the data to predict the other 30%, the AUC is very high (0.953).