Logistic regression analysis of crab data

Ma Xiaoqi

2020/5/30

1.Introduction

The natural phenomenon about horseshoe crab fertilization described by Brockmann attracts researchers'attention.The male crabs are composed of the attached and the unattached.Each female crab had a male crab attached to her in her nest.However,some female crabs have unattached male crabs as satellites around her to compete with attached male for fertilization and others have not. In our study,we need to investigate factors that affect whether the female crab had any other satellite males residing nearby her. According to the dataset of horseshoe crab,there are 173 samples on 7 columns.Each row of dataset means a female horseshoe crab and columns represent biological characteristics of female horseshoes.The seven columns are as following:

- sequnence: number of female horseshoe crab.
- weight:the weight of the female horseshoe crab,which measured in grams.
- width:carapace width of female horseshoe crab in centimeters.
- color:ordinal value range from 1 to 4,which represent light medium,medium,dark medium and dark respectively.
- spine:ordinal value range from 1 to 3,wchich stands for spine condition.The conditions given in Agresti are "both good", "one worn or broken", and "both worn or broken".
- sat:number of satellites, which means the number of males clustering around the female in addition to the male with which she is breeding.
- y:shorthand for the number of satellites.It is 1 if sat > 0.

Considering removing irrelevant variables,we finally retain 4 variables:weight,width,color and spine and labels:y.Then,we try to predict the probability that female horseshoe crab has satellites with logistic regression thus explore the most significant factors which affect the satellite condition of female crabs.

2.Exploratory Data Analysis

Firstly,we need to explore the relationshop between each of variable and label independently.For There are two kinds of variables in our study.

- categorical variables:color and spine. The boxplot in Figure 1 which form two groups based on value of y show the frequency of distribution about color and spine.From the plot,the female horseshoe crab which has the medium color and spine condition of worn and broken has the largest probability of having satellite males.
- continuous variables:weight and width.The density plot in Figure 2 show the ditribution of the numeric variable and we classify the variable into two groups based on the value of y,so we can compare the density distribution of two groups.According to the mean value of each group,on average,we can observe the female horseshoe crab which has satellites have bigger weight and width than those don't have.

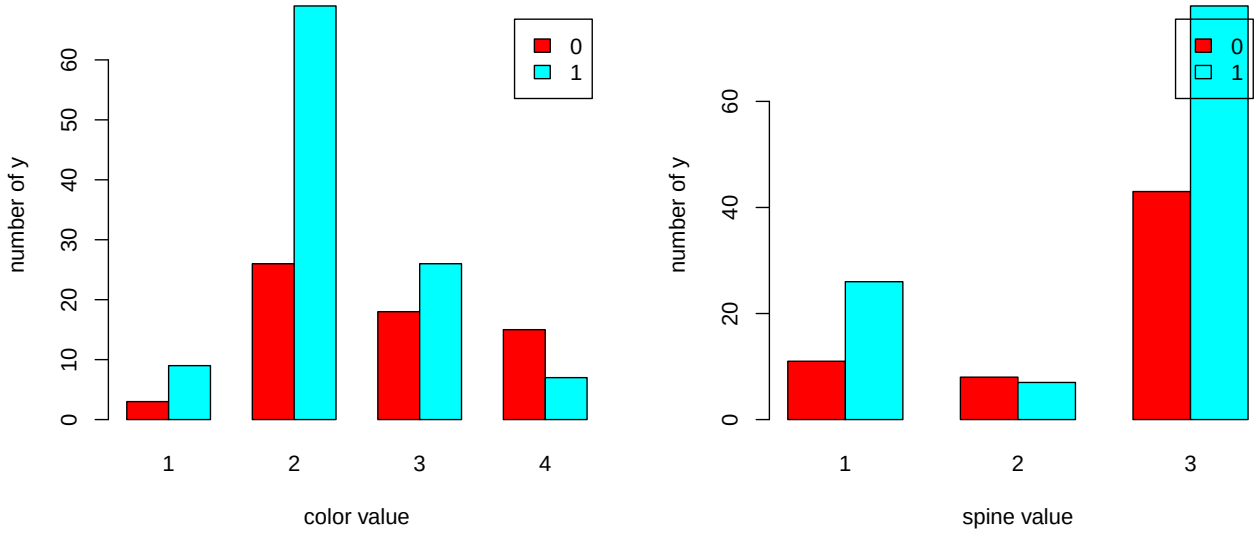

Figure 1: boxplot of categorical variables

Secondly,we try to investigate the relationship between variables.In Figure 3,we show the scatter plots and corrlation coefficients for each pair of variables.

The correlation matrix indicates a strong correlation between weight and width with correlation coefficient is 0.89.In addition,the histograms show the distribution of variables in dataset.The value of weight mainly exists in 2~3(kg) and the value of width mainly certered at interval from 22(cm) to 30(cm).As for the categorical variables,the number of medium color (value equals 2) is the most and most female horseshoe crabs has the spine condition of both worn and broken(value equals 3)

3.Baseline model

Logistic regression is our first choice to do this binary classification task.The model takes the explatanory variables x_k as input and predict the probability as output.The logit trans-

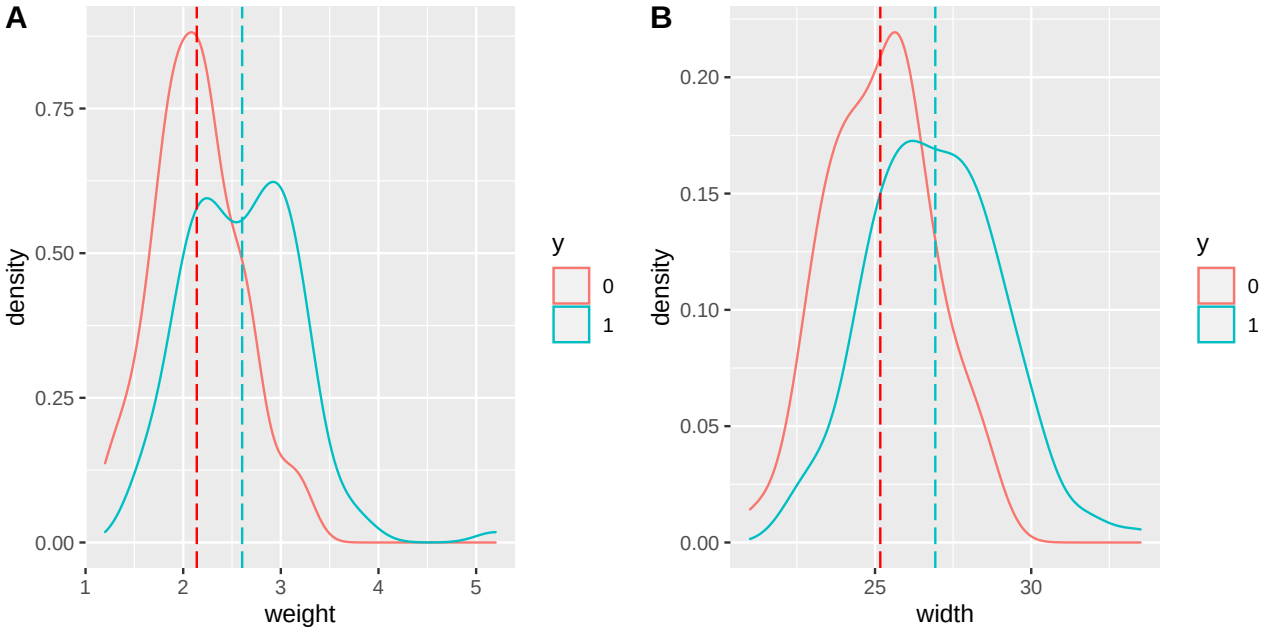

Figure 2: density plot of continuous value

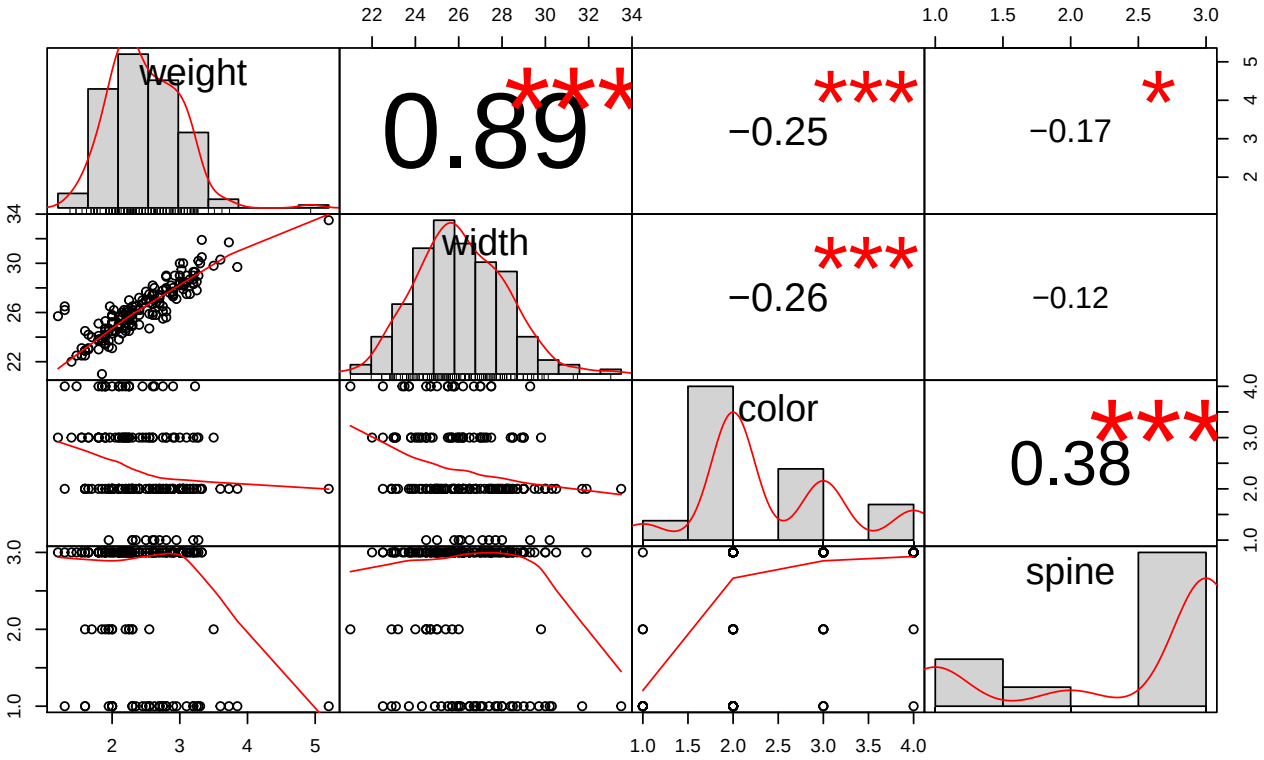

Figure 3: correlation matrix and histogrm

formation guarantee the value of p staying the interval[0,1].The coefficient of β_k represents the log-odds of variable x_k.The logistic regression model is fitted by maximum (log) likelihood.Specifically,which means minimizes the cross entropy loss in our model.

$$\text{logit}(p(\boldsymbol{x};\boldsymbol{\beta})) = \ln\left(\frac{p(\boldsymbol{x};\boldsymbol{\beta})}{1-p(\boldsymbol{x};\boldsymbol{\beta})}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m = \boldsymbol{\beta}^\top \boldsymbol{x};$$

$$p(\boldsymbol{x};\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{\beta}^\top \boldsymbol{x})}{1+\exp(\boldsymbol{\beta}^\top \boldsymbol{x})}.$$

The basic model we firstly fitted is combined with all variables in the dataset and considering the categorical values of spine and color,we transform them into factors:

$$\text{logit}(p(\boldsymbol{x};\boldsymbol{\beta})) = \ln\left(\frac{p(\boldsymbol{x};\boldsymbol{\beta})}{1-p(\boldsymbol{x};\boldsymbol{\beta})}\right) = \beta_0 + \beta_1 x_1 + \beta_1 x_2 + \vec{\beta}_3 \vec{x}_3 + \vec{\beta}_4 \vec{x}_4;$$

where $x_1 - x_2$ indicate **weight** and **width** and $x_3 - x_4$ represent the categorical vectors after factor operation and β indicate the corresponding coefficients. After minimizing the cross-entropy loss of parameters,we get the summary of the baseline model in Table 1.

Table 1: summary of baseline model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.07	3.93	-2.05	0.04
width	0.26	0.20	1.35	0.18
weight	0.83	0.70	1.17	0.24
color2	-0.10	0.78	-0.13	0.90
color3	-0.49	0.85	-0.57	0.57
color4	-1.61	0.94	-1.72	0.08
spine2	-0.10	0.70	-0.14	0.89
spine3	0.40	0.50	0.80	0.43

Table 1 shows the summary fiting results of our baseline model.In table 1, estimated represents the coefficients of variables and z-value indicates the significance of variables which shows $color_4$ is the most importmant predicator in baseline model.

From the whole perspective, we can get the residual deviance and AIC value which are defined as follows to evaluate the goodness of fit for the whole model:

- residual deviance:indicates the response predicted by a model on adding independent variables.
- Akaike Information Criterion(AIC):an estimator of out-of-sample prediction error and the relative quality of statistic model.it weighs the complexity of the estimated model and the goodness of the fitted data of the model. The formula is:

  $$AIC = 2p - 2ln(L)$$

  where p represents the number of parameters and L is the maximum likelihood of data using the model.

In the baseline model,the residual deviance is 185.20 on 165 degrees of freedom and the AIC value is 201.2.

4.Model selection

Given the baseline model,we want to look for a model with the smallest AIC(defined in section 3),which indicates a good balance between model fit and model simplicity. However,there are too many nested models for us to get CIA value for each of them,so we use the stepwise selection.

The stepwise selection do the forward step and backward step to add or remove the predictors.For example,if we start with the simplest model which just includes intercept,we should use forward step to incorporate other predictors and observe the change of AIC.On the contrary,we should use the backward step to reduce variables if we start with the most complex model which includes all variables.

With the method stepAIC in library MASS,Starting with the baseline model,we get the selected model summarized in Table 2 and the model expression is:

$$\text{logit(Y)} = log\left(\frac{Y}{1-Y}\right) = -11.39 + 0.47 * width - 0.07 * color_2 - 0.22 * color_3 - 1.33 * color_4$$

Table 2: summary of selected model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-11.385	2.873	-3.962	<0.001
width	0.468	0.106	-0.152	0.8803
color2	0.072	0.740	-0.102	0.9195
color3	-0.224	0.777	-0.233	0.8176
color4	-1.330	0.853	-0.074	0.9417

Compared with the baseline model,the selected model just retain the variabel **width** ,$color_2$,$color_3$ and $color_4$.We can find the **weight** and **spine** condition are removed by stepwise selection and $color_1$ is the default class when other three colors eqaul to 0.Futhermore,The AIC of selected model is 197.46 and its residual deviance is 187.46 on 168 degrees.

The Figure 4 shows relationship between the probability of having satellites and the variables in our selected model.From the plot,we can find that with the same value of width,the darker of the color,the lower of the probability of crabs having satellites.

5.Model assessment

5.1 Goodness of fit

5.1.1 model comparision

Based on the baseline model and selected model shown above, we compare relative parameters to evaluate the goodness of fit for two models.

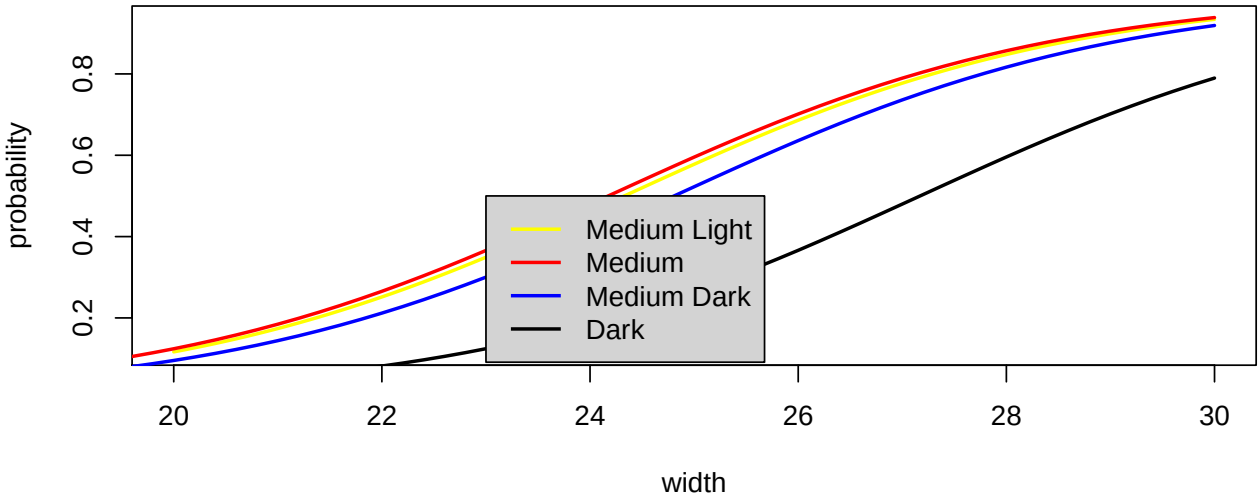

Figure 4: probability with width and color

Firstly,the baseline model which incorporates all variables has the AIC 201.2 with 185.20 on 165 degrees of freedom.The selected model which reduces variables has the AIC 197.46 with 187.46 om 168 degrees of freedom.The lower AIC value indicates better fit of data,so we think the selectd model has better fit capacity than baseline model.

Secondly,the McFadden's pseudo-R^2 is defined as $R^2 = 1 - L/L_0$,where L and L_0 represent the likelihood of fitted and null models.the McFadden's pseudo-R^2 of selected model and baseline model are 0.1697 and 0.1796.The difference between them is very small so we hope to do further test such as Likelihood ratio test to observe difference of two models.

5.1.2 Likelihood ratio test

The **likelihood ratio test** is used for comparing the goodness of fit of baseline model and selected model.It can help us test if there exists significant difference between two models. The **likelihood ratio** is defined as

$$LR = 2 * (lnL1 - lnL2),$$

where $L1$,$L2$ represent the log likelihood of selected model and baseline model.The test statistics should approximate a chi-squared random variable. The **null hypothesis** holds that the selected model is "best" and we get the p-value is 0.5212.Therefore, there is no obvious evidence for us to reject the null hypothesis.By Likelihood ration test,we find our selected model is better than baseline model.

5.2 Multicollinearity

To check for any multicollinearity between variables, we calculate the Variance Inflation Factor(VIF),which measures the effect of a set of explanatory variables on the variance of the coeffcients of another variable.Therefore,in the selected model, there is no obvious linear correlation between width and color. In the baseline model, the width (VIF:3.49)and

weight(VIF:3.39) show a linear correlation which also can be reflected by covariance matrix in section 2.

5.3 Influential observations

Considering some special points which will cause great change in our model fit,we need to do influential observations.

Firstly,we test the outliers of our model,which show there is no Studentized residuals with Bonferroni p < 0.05.Then,we calculate the cook's distance to detect the outliers which influenced the model very significantly.The points with cook's distance are greater than $4/(n-k-2)$can be regarded as outliers,where n means the number of observations and k means the number of variables.Therefore,the point 7,128,171 are regarded as influential outliers according to the cook's distance shown in Figure 5.

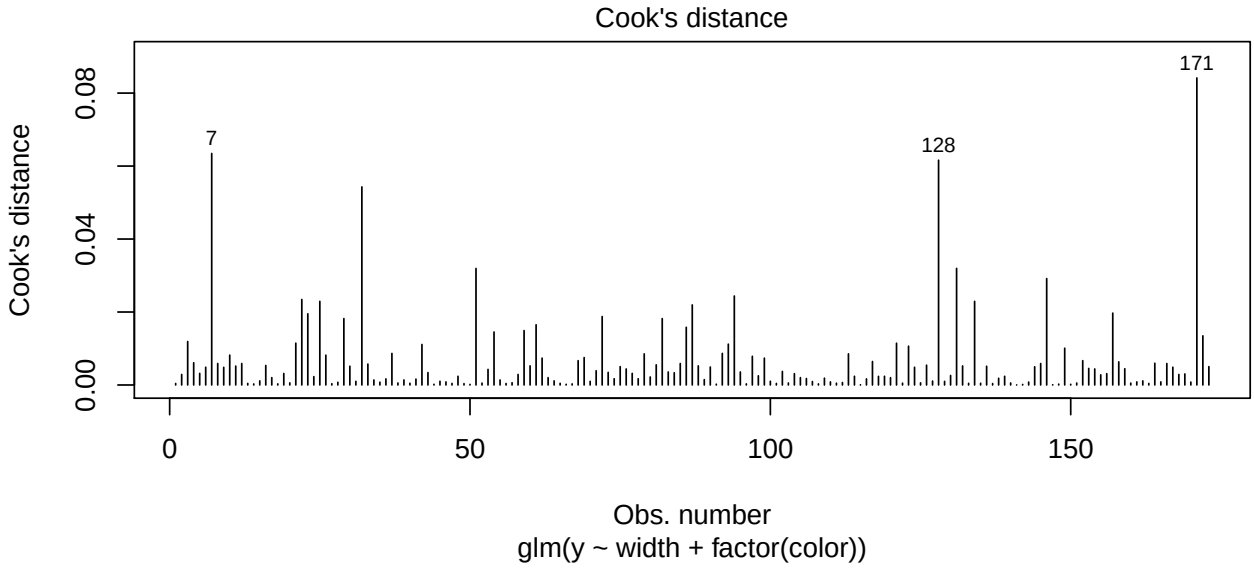

Figure 5: cook distance

Also,we implement the influential observations by influencePlot,which can show the value of residual deviance,hat statistics and cook distance in Table 3.The hat values measure how far a predicator value is different to the rest of predictor values,and the studentized residuals are calculated by fitting a model without the case for which the residual is calculated.Although we have three kinds of parameters,we still just consider the outliers calculated by cook's distance because the cook's distance shows how much the model will change if one obervation is removed.

After removing the influential outliers 7,128,171,we get the AIC value is 184.27,which shows a better fit than our selected model.

Table 3: Influence Plot

	Studres	Hat	CookD
7	-1.71	0.09	0.06
22	1.09	0.12	0.0.02
94	-2.14	0.01	0.02
128	2.20	0.03	0.06
131	1.20	0.13	0.03
171	-2.03	0.06	0.08

5.4 Cross validation

k-fold cross validation: The dataset is divided into k random folds.Each time of learning,we choose one fold as test dataset and others as training dataset.Therefore,there are k models need to be fitted.

To validate the model not overfiting,we use the 10-fold cross validation with the same model architecture of selected model after removing the influential observations.Because there is no test data set at first,so we use the 10-fold cross validation to find the optimal split of training dataset and test dataset.With the fixed seed we find the optimal training set which give the test accuracy 0.8823.Therefore,we use this training dataset to do logistic regression again and get the summary in Table 4.

Table 4: summary of cross-validation model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-11.63	3.20	-3.634	<0.001
width	0.51	0.11	4.44	<0.001
color2	-0.73	1.13	-0.64	0.52
color3	-0.88	1.16	-0.75	0.45
color4	-2.09	1.22	-1.71	0.08

5.5 Model Performance on ROC curve

In order to evaluate the performance of our final model, we plot the Receiver Operating Characteristic(ROC) which analyzes true positive,true negative,false positive,false negative.The x-axis represents the value of 1-specificity and y-axis represents the sensitivity.The specificity and sensitivity are defined as follows:

$$sensitivity = \frac{TP}{FN + TP}$$

$$specificity = \frac{TN}{TN + FP}$$

Therefore,the value of Area Under the Curve(AUC) in whole dataset(without influential outliers) is 0.794.The figure 6 shows below is the ROC curve in the whole dataset with the coefficients from optimal training dataset choosed by 10-fold cross-validation.

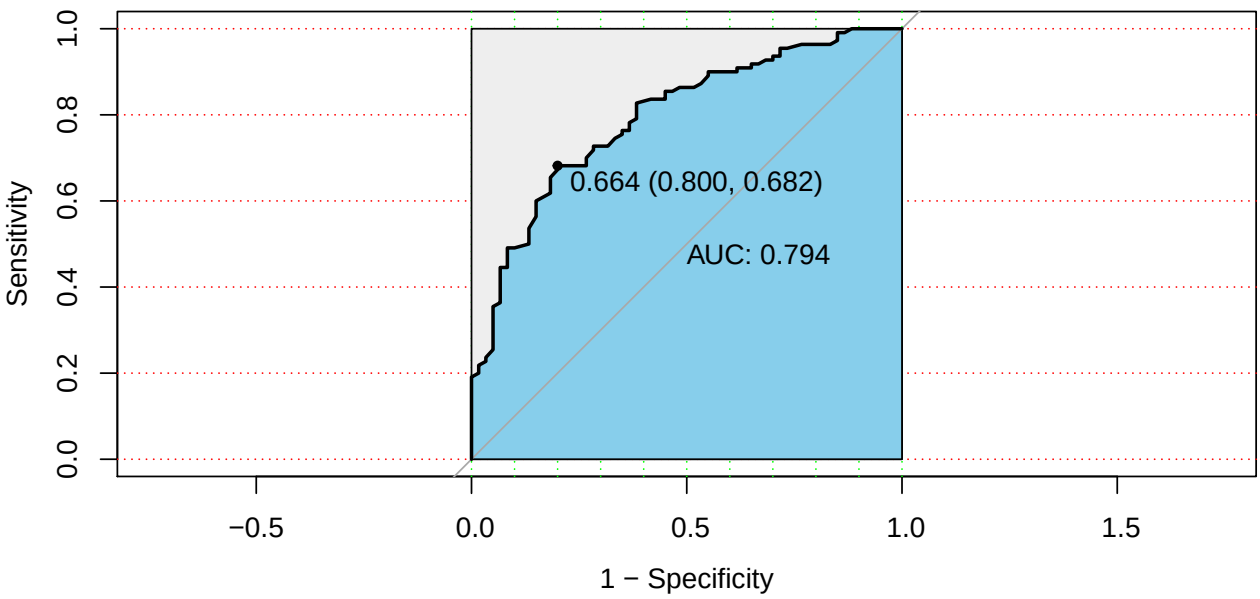

Figure 6: ROC curve

6.Conclusion and discussion

In this project, we use a logistc regression model to analyze the factors which affects the female satellites condition. Starting with the baseline model, we do the model selection in terms of AIC value to choose the model with better fitness. After deleting the influential observations by cook's distance, we do the 10-fold cross validation on the new dataset and choose the parameters which achieve best performance.

Our final estimated model is:

$$\text{logit(Y)} = log\left(\frac{Y}{1-Y}\right) = -11.63 + 0.51 * width - 0.73 * color_2 - 0.88 * color_3 - 2.09 * color_4$$

In our model, we select the two variables: **width** and the **color** as the last predicators to do classification. Finally, we get the model with AUC value of 0.794 on the whole dataset,which shows a good performance.The model removes the variable **weight** which has a strong correlation with **width**, also **spine** which is not significant when predicting the female satellite conditions.At the same time,we can see that **width** has a positive relationship for the female crab having satellites which means the wider female crab is,the higher probability of female crab have satellites. **color** has a negative relationship for female crab having satellites. Since **color** is a kind of categorical variable, we can find the darker of the color will cause the lower of the probability of crabs having satellites.

Reference

Brockmann H J. Satellite male groups in horseshoe crabs, Limulus polyphemus[J]. Ethology, 1996, 102(1): 1-21.