24 Hour Project-II

Yuanyou Yao

December 9, 2019

Contents

1 Introduction

The purpose of the article is to model the survival of male sparrows (1 = survived, 2 = perished) and physical characteristics of the birds. Belows are explanatory:

1. AG : Ages (1 for adults, 2 for juveniles);
2. TL : Total length (mm);
3. AE : Alar extent (mm);
4. WT : Weight (g);
5. BH : Length of beak and head (mm);
6. HL : Length of humerus (inch);
7. FL : Length of femur (inch);
8. TT : Length of tibio-tarsus (inch);
9. SK : Width of skull (inch);
10. KL : Length of keel of sternum (inch).

We firstly treat the survival status as *Bernoulli Distribution* and then use *Generalized Linear Regression(GLM)* to find their relationship. In order to find top two fit models, we delete some interactions and use *Stepwise* and the final models are selected by AIC and BIC.

After two models are obtained, we perform two test—*Wald Test* and *Likelihood Ratio Test(LRT)*—to verify. More importantly, we are interested in predictivability of models. Therefore, a *repeated random sub-sampling validation* is performed and conclusion can be drawn.

Accordingly, all R codes are put in the appendix.

2 Methods and Models

2.1 Preparation

2.1.1 For Response Variable

Since survival has two status—*Survived* and *Perished*, we use 1 = survived, 2 = perished to show (But in R, we would like to use 0 = perished to satisfy *Logistic Regression Model*).

Then we use π_i to denote the success probability of Y_i

2.1.2 For Explanatory Variable

There are 55 variables including original ones and interactions. However, because some interactions are difficult to interpret, we decide to drop them, which are related to the variables KL, SK, TT, FL and HL.

Then, we can do *Logistic Regression*.

2.2 All Variables Regression

2.2.1 Coeffiecients

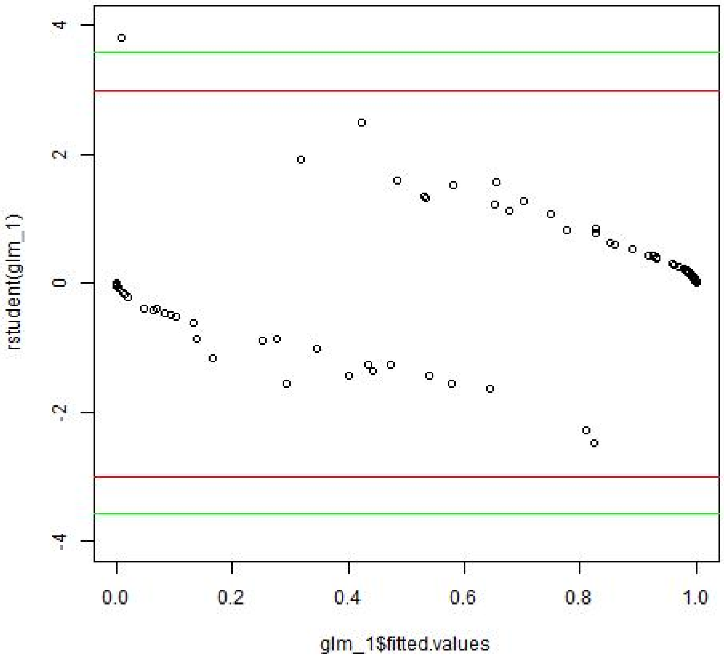
We simply do a logistic regression with all variables and get the answer:

```
Coefficients:
             Estimate
(Intercept) 3617.1184
AG          -138.1104
TL           -30.0234
AE             5.6430
WT           -28.1808
BH           -95.2689
HL            93.3188
FL            49.3599
```

TT	−21.0532
SK	32.1328
KL	39.1353
‘AG:TL‘	0.4236
‘AG:AE‘	−0.6298
‘AG:WT‘	0.6994
‘AG:BH‘	6.5736
‘TL:AE‘	0.0336
‘TL:WT‘	−0.4509
‘TL:BH‘	0.9992
‘AE:WT‘	0.1794
‘AE:BH‘	−0.4686
‘WT:BH‘	1.7102

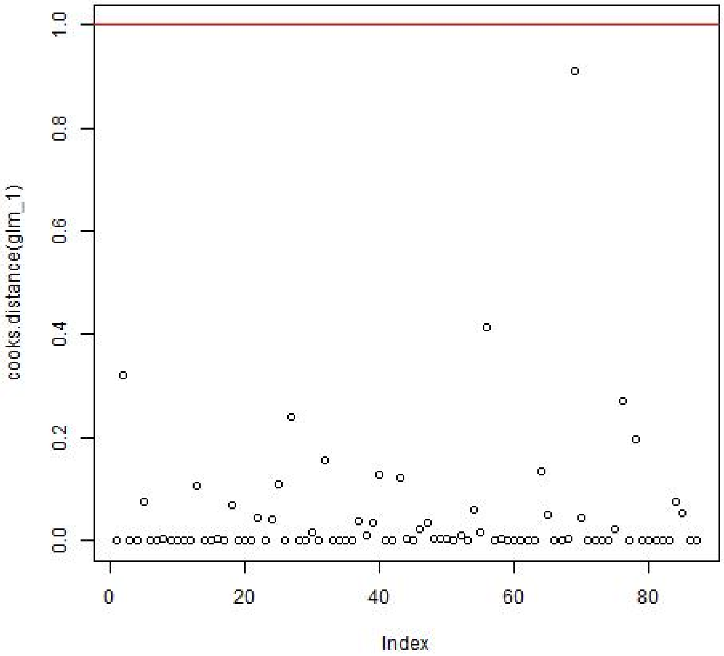
2.2.2 Outliers and Influntial Observations

We initially check whether the model has any outliers or influntial observations. To begin with, we check studentized residuals and get the following graph:



By rule of thumb: $|t_i| > t_{n-p-1,1-\frac{\alpha}{2n}}$, the figure tells us that $NO.3\ Y$ is an outlier. So, it should be deleted.

Besides, we check influntial observations using *cook's distance* and get the following graph:



Based on the image,we can say that there are no influntial observations.

2.3 Stepwise Selection: Forward Model

2.3.1 Model

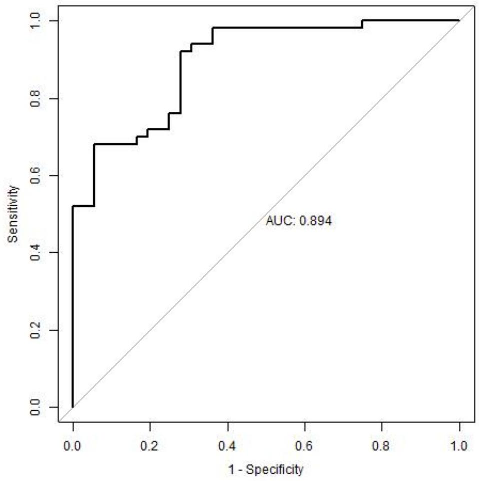
Based on the model in 2.2 and 3^{rd} observation has been dropped, forward selection is applied to find best model. According to R, we get two models based on AIC and BIC.

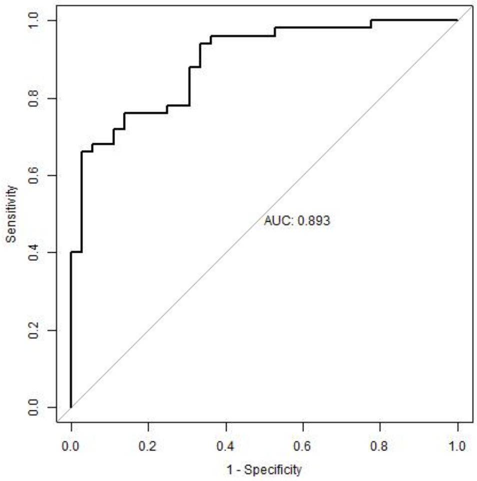
$$AIC : Status = 53.54683 - 0.71944TL + 58.16916HL - 0.90114WT + 23.76916KL + 0.00291AE * BH$$

$$BIC : Status = 49.1871 - 0.6492TL + 72.0453HL - 0.7942WT + 27.1529KL$$

2.3.2 Tests

Then we consider AUCs:





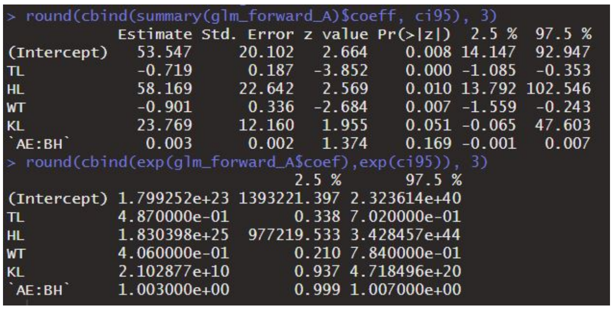
Thus, we choose the first model as one of the best models $i.e.$

$$Status = 53.54683 - 0.71944TL + 58.16916HL - 0.90114WT + 23.76916KL +$$

$$0.00291AE * BH$$

Generally, AUC is 0.894 and we can interpret it as excellent discrimination

In addition, *Wald Test* and *Likelihood Ratio Test(LRT)* are performed to ensure the model is valid. Belows are some of the result from R:



```
> round(cbind(summary(glm_forward_A)$coeff, ci95), 3)
            Estimate Std. Error z value Pr(>|z|)  2.5 %  97.5 %
(Intercept)   53.547     20.102   2.664    0.008 14.147  92.947
TL            -0.719      0.187  -3.852    0.000 -1.085  -0.353
HL            58.169     22.642   2.569    0.010 13.792 102.546
WT            -0.901      0.336  -2.684    0.007 -1.559  -0.243
KL            23.769     12.160   1.955    0.051 -0.065  47.603
`AE:BH`        0.003      0.002   1.374    0.169 -0.001   0.007
> round(cbind(exp(glm_forward_A$coef),exp(ci95)), 3)
                              2.5 %       97.5 %
(Intercept) 1.799252e+23 1393221.397 2.323614e+40
TL          4.870000e-01       0.338 7.020000e-01
HL          1.830398e+25  977219.533 3.428457e+44
WT          4.060000e-01       0.210 7.840000e-01
KL          2.102877e+10       0.937 4.718496e+20
`AE:BH`     1.003000e+00       0.999 1.007000e+00
```

```
> Anova(glm_forward_A, type="III")
Analysis of Deviance Table (Type III tests)

Response: y
       LR Chisq Df Pr(>Chisq)
TL      25.6980  1  3.992e-07 ***
```

```
HL          8.1424   1    0.004324 **
WT          9.5551   1    0.001994 **
KL          4.4056   1    0.035822 *
‘AE:BH‘     2.0003   1    0.157264
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 “. 0.1 “  1
```

As we can see, the model selected above satisfies the tests well.

2.3.3 Interpretation

We consider the first observation, which is:

Intercept	TL	HL	WT	KL	$AE * BH$
1	154	0.69	24.5	0.83	7519.2

Then we get the success probability

$$\pi_1 = \frac{exp(X^T\hat{\beta})}{1 + exp(X^T\hat{\beta})} = 0.7147148$$

When these variables are fixed, the probability that the sparrows survive is $\pi_1 = 0.7147148$

We think about the coefficients of TL, which $\beta_1 = -0.71944$. Consider $exp(\beta_1) = 0.4870226$ and it is called *odd ratio*. It means that given other varibles fixed, when total length be 1 mm longer, sparrows are more likely to die during the winter storm. The probability they will survive will be $0.4870226 * 0.7147148 = 0.3480823$.

2.4 Stepwise Selection: Backward Model

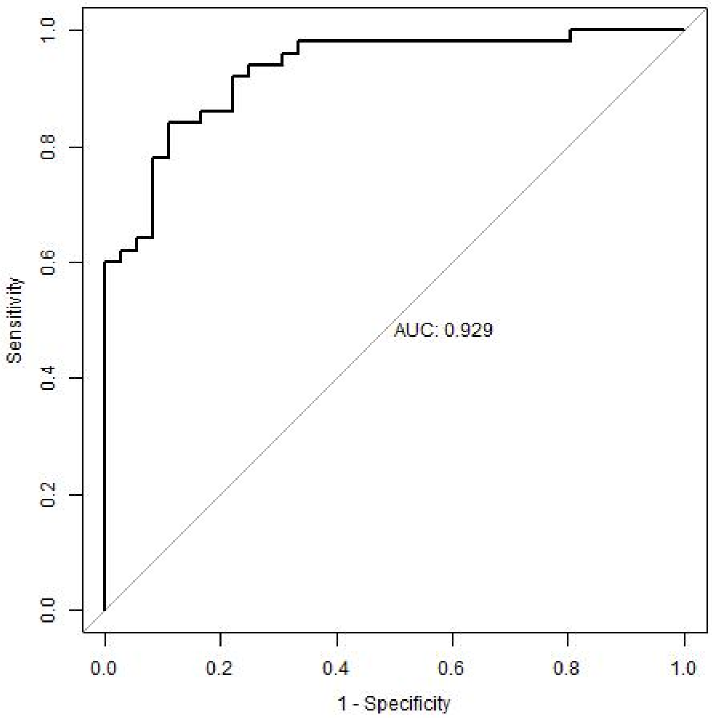
2.4.1 Model

Based on the model in 2.2 and 3^{rd} observation has been dropped, backward selection is applied to find best model. According to R, both AIC and BIC find the same model which is:

$$Status = 165.323559 - 1.649037TL - 3.814491BH + 72.846849HL + 34.281506KL - 0.458239AG * AE + 3.594124AG * BH + 0.004119TL * AE - 0.006282TL * WT$$

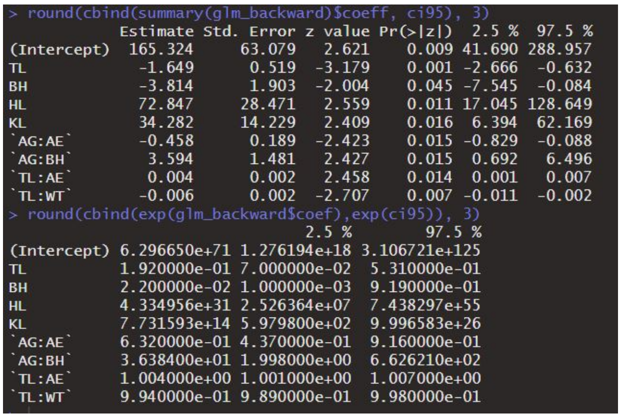
2.4.2 Tests

At first we check AUC:



Generally, AUC is 0.929 and we can interpret it as outstanding discrimination.

In addition, *Wald Test* and *Likelihood Ratio Test(LRT)* are performed to ensure the model is valid. Belows are some of the result from R:



```
> round(cbind(summary(glm_backward)$coeff, ci95), 3)
            Estimate Std. Error z value Pr(>|z|)  2.5 %  97.5 %
(Intercept)  165.324     63.079   2.621    0.009 41.690 288.957
TL            -1.649      0.519  -3.179    0.001 -2.666  -0.632
BH            -3.814      1.903  -2.004    0.045 -7.545  -0.084
HL            72.847     28.471   2.559    0.011 17.045 128.649
KL            34.282     14.229   2.409    0.016  6.394  62.169
`AG:AE`       -0.458      0.189  -2.423    0.015 -0.829  -0.088
`AG:BH`        3.594      1.481   2.427    0.015  0.692   6.496
`TL:AE`        0.004      0.002   2.458    0.014  0.001   0.007
`TL:WT`       -0.006      0.002  -2.707    0.007 -0.011  -0.002
> round(cbind(exp(glm_backward$coef),exp(ci95)), 3)
                             2.5 %        97.5 %
(Intercept) 6.296650e+71 1.276194e+18 3.106721e+125
TL          1.920000e-01 7.000000e-02  5.310000e-01
BH          2.200000e-02 1.000000e-03  9.190000e-01
HL          4.334956e+31 2.526364e+07  7.438297e+55
KL          7.731593e+14 5.979800e+02  9.996583e+26
`AG:AE`     6.320000e-01 4.370000e-01  9.160000e-01
`AG:BH`     3.638400e+01 1.998000e+00  6.626210e+02
`TL:AE`     1.004000e+00 1.001000e+00  1.007000e+00
`TL:WT`     9.940000e-01 9.890000e-01  9.980000e-01
```

```
> Anova(glm_backward, type="III")
Analysis of Deviance Table (Type III tests)

Response: y
```

```
           LR Chisq  Df  Pr(>Chisq)
TL          14.9124   1   0.0001126 ***
BH           5.1889   1   0.0227317 *
HL           8.1407   1   0.0043283 **
KL           7.3669   1   0.0066434 **
‘AG:AE‘      8.9230   1   0.0028160 **
‘AG:BH‘      8.9525   1   0.0027709 **
‘TL:AE‘      8.1827   1   0.0042291 **
‘TL:WT‘     10.3317   1   0.0013076 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ”. 0.1 “  1
```

As we can see, the model selected above satisfies the tests well.

2.4.3 Interpretation

We consider the first observation, which is:

Intercept	TL	GH	HL	KL	$AG*AE$	$AG*BH$	$TL*AE$	$TL*WT$
1	154	31.2	0.69	0.83	241	31.2	37114	3773

Then we get the success probability

$$\pi_1 = \frac{exp(X^T\hat{\beta})}{1+exp(X^T\hat{\beta})} = 0.7055035$$

When these variables are fixed, the probability that the sparrows survive is $\pi_1 = 0.7055035$

We think about the coefficients of TL, which $\beta_1 = -1.649037$. Consider $exp(\beta_1) = 0.1922349$ and it is called *odd ratio*. It means that given other varibles fixed, when total length be 1 mm longer, sparrows are more likely to die during the winter storm. The probability they will survive will be $0.1922349 * 0.7055035 = 0.1356224$.

3 Results

To campare the predictability of two models above, *cross validation* can be applied. Belows are the main steps:

1. Split the data into 2 parts and name as train and test. The proportion used of train and test sample is .75.

2. Use the explanatory variables in 2.3 and 2.4 to do GLM with trained data, and use test data for prediction.

3. Record the 2 Area Under the ROC Curve (AUC).

4. Repeat (1) to (3) for 500 times. Calculate the mean AUCs and compare the two numbers.

Here is what we get:

$$AUC.forward = 0.8462902$$

$$AUC.backward = 0.8822151$$

Consequently, we get the best model:

$$Status = 165.323559 - 1.649037TL - 3.814491BH + 72.846849HL + 34.281506KL - 0.458239AG * AE + 3.594124AG * BH + 0.004119TL * AE - 0.006282TL * WT$$

4 Limitations

1. There must be some useful interactions for us to predict but we arbitrarily delete them. For example, $TT * SK$ means the product of tibio-tarsus and skull length, which could an significant level to evaluate the stamina of sparrows to find food in a storm and contribute them to survive. As a consequence, our model cannot reflect all aspects that survival is related.

2. As for original data, we check outliers and influntial observations. It is neccessary to notice that we use *standardized residuals* and *student residuals* to check outliers and *DFFITS*, *cook's distance* and *DFBETAS* to check

observation. However, these methods give different results that will be attached in appendix. In our analysis, we just use one to decide whether the obsevations should be drop or not. If we use other methods, which will delete more according to calculation, the final model can be more convincing.

3. It can be infer from 2.3.2, while doing LRT, the p-value of variable $AE*BH$ is 0.157264.So it is not as good as the model in 2.4.1. Correspondently, the predictability support our assertion.

5 Conclusions

In the report, we propose top two models that is used to find how survival is related to physical characteristics. Initially, we use the whole data to do *logistic regression* and check the validation of data. After deleting some influtial observations, we then conduct *stepwise selection* to reduce the variables. Finally, we get two models that satisfy our requirement by *wald test* and*LRT*. A *cross validation* is performed to reach the goal of predictability and we have the best model:

$$Status = 165.323559 - 1.649037TL - 3.814491BH + 72.846849HL + 34.281506KL -$$

$$0.458239AG*AE + 3.594124AG*BH + 0.004119TL*AE - 0.006282TL*WT$$

When choose the best-fit models, we give some interpretations about coefficients and some details concerning to tests have been mentioned as well.

We also look back the whole analysis and find some critical issuses like dropped variables that can change the final models to a considerable degree.

A Appendix R Code

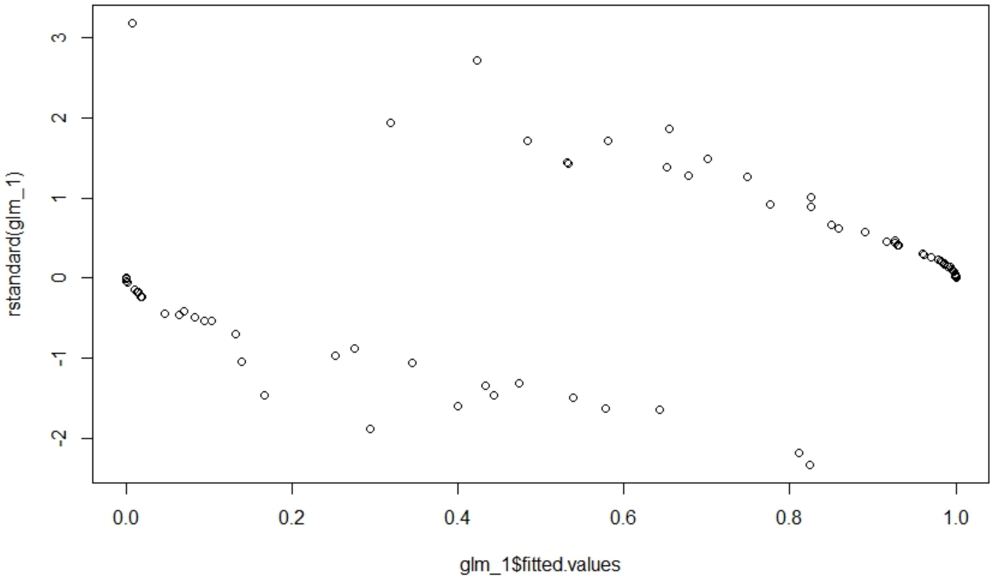
```
library(car)
library(bestglm)
```

```
data=read.csv(”24H_project2.csv”,header = T,
stringsAsFactors = F)
data$Status[data$Status==”Survived”]=1
data$Status[data$Status==”Perished”]=0
data$Status=as.numeric(data$Status)
table(data$Status)
prop.table(table(data$Status))
f <- as.formula(y ~ .*.)
y <- data$Status
x <- model.matrix(f, data[,-1])[, -1]
x=x[,-c(54,53,50,47,43,38,32,25,17,55,54,52,51,48,49,45,
44,40,39,34,33,26,27,19,18,15,16,23,24,30,31,37,36,41,42,
46,47,50)]
data_new=as.data.frame(cbind(y,x))
attach(data_new)
glm_1=glm(y~.,data = data_new,family = binomial(”logit”))
summary(glm_1)
jpeg(”glm_1.jpg”)
par(mfrow=c(2,2), pty=”s”)
plot(glm_1)
dev.off()
cbind(resid(glm_1),rstandard(glm_1),rstudent(glm_1))
plot(glm_1, which=3)
plot(glm_1$fitted.values,sqrt(abs(rstandard(glm_1))),
ylim=c(0,1.5))
plot(glm_1$fitted.values,rstandard(glm_1))
# check studentized residuals
jpeg(”rstudent.jpg”)
plot(glm_1$fitted.values,rstudent(glm_1), ylim=c(-4,4))
abline(h=c(-3,3), col=”red”)
# Bonferroni correction
alpha <- 0.05
```

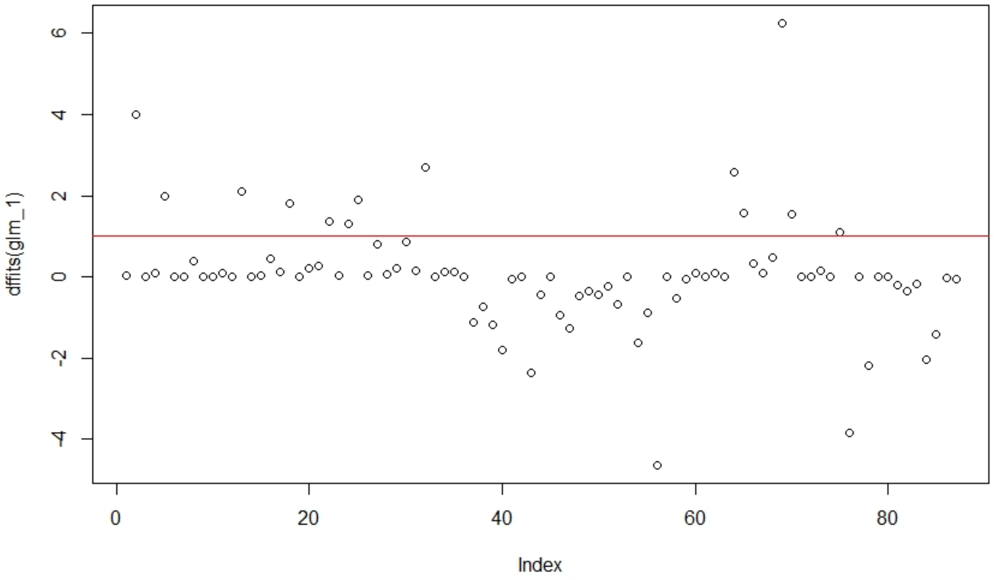
```
t.critical <-qt(1-alpha/(2*n), n-p-1)
abline(h=c(-t.critical, t.critical), col=”green”)
dev.off()
glm_1$fitted.values[max(rstudent(glm_1))]
# Identifying Influential Observations#
dffits(glm_1)
plot(dffits(glm_1))
abline(h=1, col=”red”)
# Cook’s distance
cooks.distance(glm_1)
plot(glm_1, which = 4)
jpeg(”cook.jpg”)
plot(cooks.distance(glm_1),ylim = c(0,1))
abline(h=1, col=”red”)
dev.off()
# DFBETAS
dfbetas(glm_1)
plot(dfbetas(glm_1)[,2])# DFBETAS_{1(i)}
abline(h=1, col=”red”)
data_new=data_new[-3,]
y=data_new$y
glm_1=glm(y~.,data = data_new,family = binomial(”logit”))
step(glm_1,k=log(length(y)))#backward+BIC
step(glm_1)#backward+AIC
if(!require(”pROC”)) {
  install.packages(”pROC”)
  stopifnot(require(”pROC”))}
glm_backward=glm(formula = y ~ TL + BH + HL + KL + ‘AG:AE‘ +
  ‘AG:BH‘ + ‘TL:AE‘ + ‘TL:WT‘, family = binomial(”logit”),
  data = data_new)
backward.pROC <-roc(y~ fitted(glm_backward))
jpeg(”backward.jpg”)
```

```
plot.roc(backward.pROC, legacy.axes=TRUE, print.auc=TRUE)
dev.off()
ci95 <-confint.default(glm_backward)
round(cbind(summary(glm_backward)$coeff, ci95), 3)
round(cbind(exp(glm_backward$coef),exp(ci95)), 3)
Anova(glm_backward, type=”III”)
attach(data_new)
glm0 <-glm(y~1, family=binomial(”logit”))
step(glm0, scope =list(upper=glm_1),
direction = ”forward”)#forward+AIC
step(glm0,scope = list(upper=glm_1),direction = ”forward”,
k=log(length(y)))#forward+BIC
glm_forward_A=glm(formula = y ~ TL + HL + WT + KL+‘AE:BH‘,
family = binomial(”logit”), data = data_new)
glm_forward_B=glm(formula = y ~ TL + HL + WT + KL,
family = binomial(”logit”), data = data_new)
forward.pROC.A <-roc(y~ fitted(glm_forward_A))
forward.pROC.B <-roc(y~ fitted(glm_forward_B))
jpeg(”forwardA.jpg”)
plot.roc(forward.pROC.A, legacy.axes=TRUE, print.auc=TRUE)
dev.off()
jpeg(”forwardB.jpg”)
plot.roc(forward.pROC.B, legacy.axes=TRUE, print.auc=TRUE)
dev.off()
ci95 <-confint.default(glm_forward_A)
round(cbind(summary(glm_forward_A)$coeff, ci95), 3)
round(cbind(exp(glm_forward_A$coef),exp(ci95)), 3)
Anova(glm_forward_A, type=”III”)
model1.auc=vector()
model2.auc=vector()
for (i in 1:500) {
p <- 0.75 # ratio btw train vs. valid, which you can decide
idx <-sample.int(n = nrow(data_new),
 size = floor(p*nrow(data_new)),replace = FALSE)
train_data <- data_new[idx,]
test_data <- data_new[-idx,]
#Fit the two models with train data.

model1.trained <- glm(formula = y ~ TL + HL + WT +
KL +AE*BH, family = binomial(”logit”), data=train_data)
model2.trained <- glm(formula = y ~ TL + BH + HL + KL +
 ‘AG:AE‘ + ‘AG:BH‘ + ‘TL:AE‘ +  ‘TL:WT‘,
family = binomial(”logit”), data=train_data)
#Predict your response with validation data.

model1.pred <- predict.glm(model1.trained,
 newdata = test_data, type=”response”)
model2.pred <- predict.glm(model2.trained,
 newdata = test_data, type=”response”)

#Compute the AUC and record it
#(because we will repeat this 500 times).

 if (!require(”pROC”)) {
   install.packages(”pROC”)
   stopifnot(require(”pROC”))
}
model1.auc[i] <- auc(roc(test_data$y, model1.pred))[[1]]
model2.auc[i] <- auc(roc(test_data$y, model2.pred))[[1]]
#Repeat a) to d) 500 hundred times
# and record your AUCs, and then average.
}
(model1.ave=mean(model1.auc))
(model2.ave=mean(model2.auc))
```

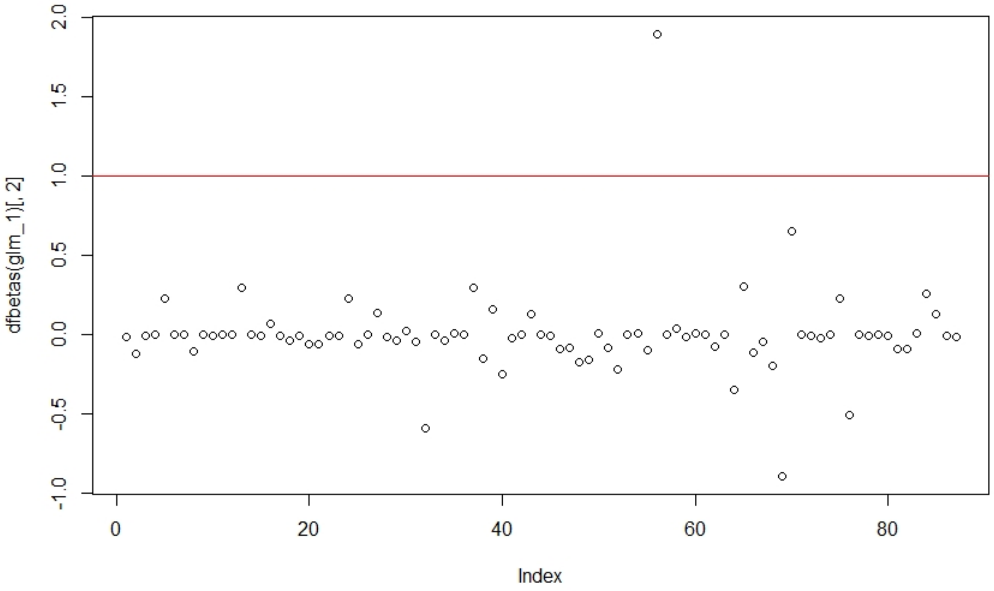
B Output of R



rstandard(glm_1)
3
2
1
0
-1
-2
0.0
0.2
0.4
0.6
0.8
1.0
glm_1$fitted.values



dffits(glm_1)
6
4
2
0
-2
-4
0
20
40
60
80
Index


dfbetas(glm_1)[, 2]
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
0
20
40
60
80
Index