# 207 PROGECT

## Survival of House Sparrows

Yi Zhou

(998086084)

YanLin Li

(998303301)

MARCH 22, 2017

# Survival of House Sparrows


## Abstract

The data recorded various physical characteristics of house sparrows which were found on the ground after a severe winter storm in 1898. Among these house sparrows, some survived and some perished. We are interested in the relationship between the probability of survival and sparrows' physical characteristics.

Our final model suggests that the survival of sparrows is found to be negative related to total length and weight, but positive related to length of humerus, length of keel of sternum, and square of femur length. We conclude sparrows with lower body weight, and larger body bone including longer humerus, femur and keel of sternum lengths tend to survive.


## Introduction

Usually, people consider sparrows with larger body size tend to survival. Some studies have found that survival rate increased significantly with greater general size, including longer humerus, femur and keel of sternum lengths and wider skulls. Zoologists also found that sparrows with lower body weight and shorter total length have more probability to survive.

In this report, we are interested in how these physical characteristics of house sparrows influence their survival status during winter, and want to see if our conclusion support or disagree with these scientific findings. Since the response variable is binary, we use logistic regression model to fit the data.

## Data Processing

The dataset is pretty clear, and does not have 'NA' or missing values. It consists two qualitative variables including the response variable, and 9 quantitative variables. Looking at the scatter plot matrix of 9 quantitative variables (Figure 1), there are no obvious outlier in the plot, but some first order multicollinearity are discerned among some variables. Figure 2 shows the multicollinearity more explicitly: We can see that alar extent, length of humerus, length of femur, and length tibio-tarsus have high multicollinearity to each other. This make sense, as larger sparrows tend to have longer length in wings and bones at the same time.

Histograms (Figure 3) are performed on the 9 quantitative variables. Most of the histograms appear to be normal, whereas the Weight variable is left skewed and the Beak Head Length is right skewed. Therefore, we perform log transformation on these two variables. Since the quantitative variables are measured in different scales, we also standardize all the quantitative variables, and plot the histogram of them again (Figure 4). All the histograms looks normal and centered at zero. We also fit a lowess fit line for response variable against the 9 quantitative variables (Figure 5). Most of the lowess line in the plots are linear, so it is reasonable to fit the initial model with first order variables.

## Model Selection

In this part, we use forward stepwise procedure based on both AIC and BIC criterion, and then fit the model with second order and interaction term.

1. **First-order model selection**

   We fit the model with all first-order effects as our model 1. Our full model is

logit(Status) = $\beta_0$ + $\beta_1$Age + $\beta_2$TotalL + $\beta_3$AlarExtL+ $\beta_4$log(Wght) + $\beta_5$log(BeakHeadL) + $\beta_6$HumerusL+ $\beta_7$FemurL+ $\beta_8$TribioTarsusL + $\beta_9$SkullWd + $\beta_{10}$KeelL and the empty model is logit(Status) = $\beta_0$. Both forward selection or backward selection with AIC and BIC criterion converges to the same model (the summary and ANOVA table of this model is in figure 6 and 7): logit(Status) = 0.6335 - 2.0292*TotalL - 1.0690*ln(Wght) + 1.6133 HumerusL + 0.9245*KeelL with AIC: 79.73.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.6335     0.3175   1.995 0.045998 *
TotalL       -2.0292     0.5231  -3.880 0.000105 ***
Wght         -1.0690     0.4192  -2.550 0.010776 *
HumerusL      1.6133     0.4709   3.426 0.000613 ***
KeelL         0.9245     0.4119   2.244 0.024820 *
---
```

Above is the ANOVA table for the resulting best model. We can use Wald Test to test single β.
Ho: $\beta_k = 0$ v.s. $H_1$: $\beta_k \neq 0$ (k = 0…4)
$Z^* = b_k/s\{ b_k \}$
All the p-values of the x variables are smaller than $\alpha = 0.05$, so we fail to reject $H_0$, indicating none can be drop individually from the model.

We also compare this model with the saturated model (ANOVA model is in the figure 8):
Ho: logit(Status)=0.6335-2.0292*TotalL-1.0690*log(Wght) +1.6133 HumerusL+0.9245*KeelL
H1: The model is not a good fit.
$G^2$ = -2[logL(R) – logL(F)] = 69.728 – 65.698 = 0.03
($G^2$: Residual Deviance of Reduced model - the Residual Deviance of the full model)
$\chi^2$ (1-0.05, 82-76) = $\chi^2$ (0.95, 6) = 12.6
Since $G^2$ is much smaller than $\chi^2$ (0.95, 6) = 12.6, we fail to reject the null hypothesis. The model logit(Status)=0.6335-2.0292*TotalL-1.0690*ln(Wght) +1.6133 HumerusL+0.9245*KeelL is a good fit of the sparrow data.

## 2. Second-order model selection

We fit the full model with all first order and second order without interaction terms:
logit(Status) = $\beta_0$ + $\beta_1$Age + $\beta_2$TotalL +…..+$\beta_{10}$KeelL+ $\beta_{11}$Total$^2$+$\beta_{11}$AlarExtL$^2$+….+$\beta_{20}$ KeelL$^2$
and the empty model is logit(Status) = $\beta_0$. Here all the x quantitative variables are set as centered $x = x - \overline{x}$ . Since the full model has lots of variables and some variables are highly correlated with each other as we explained in the Data Processing section, using forward selection will be better than backward selection in this situation. Forward selection method can deal with multicollinearity better compared with backward selection, and it will drop the highly correlated variable automatically. By using forward selection with AIC and BIC, we get the selected model as:
logit(Status) = TotalL + HumerusL + ln(Wght) + KeelL + FemurL$^2$ + BeakHeadL + HumerusL$^2$

After we have the best model chosen by forward selection method, we then set this model as a new full model, and put it into backward selection. This method can give us a better model compared with only fitting forward selection method. The final model for second-order model by both AIC and BIC is: logit(Status)=0.1472-2.2760*TotalL+ 2.0814*HumerusL -1.1443*ln(Wght) + 0.9354*KeelL + 0.5426*FemurL$^2$ with AIC = 77.22 (The summary and ANOVA table of this model is in figure 8 and 9)

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.1472     0.3990   0.369 0.712218
TotalL       -2.2760     0.5786  -3.934 8.37e-05 ***
HumerusL      2.0814     0.5658   3.679 0.000234 ***
Wght         -1.1443     0.4186  -2.733 0.006268 **
KeelL         0.9354     0.4158   2.250 0.024456 *
I(FemurL^2)   0.5426     0.2826   1.920 0.054827 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can use Wald Test to test single β. Ho: $\beta_k = 0$ v.s. $H_1$: $\beta_k \neq 0$ (k = 0…5). Here the p-value of $\beta_1, \beta_2, \beta_3$ and $\beta_4$ are less than $\alpha = 0.05$, we fail to reject the null hypothesis. The p-value of the second order variable is a little bit higher than $\alpha = 0.05$, but it is very close to 0.05. We can further use Deviance Goodness of fit test to test if the model is appropriate.

We also compare this selected model with the saturated model.(ANOVA model of saturated model is in the figure 11)

Ho: logit(Status)=0.1472-2.2760*TotalL+2.0814*HumerusL -1.1443*ln(Wght) + 0.9354*KeelL + 0.5426*FemurL$^2$

H1: The model is not a good fit.

$G^2 = -2[\log L(R) - \log L(F)] = 65.223 - 51.814 = 13.409$

$\chi^2$ (1-0.05, 81-67) = $\chi^2$ (0.95, 14) = 23.7

Since $G^2$ is much smaller than $\chi^2$ (0.95, 14) = 23.7, we fail to reject the null hypothesis. The model is: logit(Status)=0.1472-2.2760*TotalL+2.0814*HumerusL -1.1443*ln(Wght) + 0.9354*KeelL + 0.5426*FemurL$^2$ , is a good fit of the sparrow data.

### 3. Interaction Selection

We also fit the model with all the first order and interaction term and use forward selection to choose the best model. The model chosen by forward selection procedure is exactly the same as the best model in the first-order model: logit(Status) = 0.6335 - 2.0292*TotalL - 1.0690*ln(Wght) + 1.6133 HumerusL + 0.9245*KeelL.Therefore, we selected two model to fit the data, one is in the first order and one is in the second:

First order model:

*logit(Status) =0.6335-2.0292*TotalL-1.0690*ln(Wght)+1.6133HumerusL+0.9245*KeelL*

Second order model:

*logit(Status)=0.1472-2.2760*TotalL+2.0814*HumerusL-1.1443*ln(Wght)+0.9354*KeelL+ 0.5426*FemurL$^2$*

We will further analyze these two models and selected a final model in the Model Diagnostic part.

## Model Diagnostic

We fit the two models individually and plot the Pearson residual, Studentized Pearson residuals and Deviance residuals against fitted probability, as well as half normal plot (Figure12&Figure14). We also look at the scatter plot of the residuals, and noticed possible outliers (Figure13&Figure15). We removed the corresponding outliers and refit the models again. We compare the model by looking at 95% confidence interval for all estimate coefficients, 5 fold cross validation accuracy, proportion of reduction error which we define as 1- $\frac{\sum(y_i - \hat{\pi}_i)^2}{\sum(y_i - \bar{y})^2}$, area under the curve, and its 95% confidence interval for AUC. Below is a summary of the two models:

Model 2 wins over model 1 in all aspect, so we chose model 2 to be our final logistic model:

**logit(Status) = -0.02171 - 2.70422*Total _Length - 1.41988* ln(Wight) + 2.59218*Humerus_Length + 1.27886*Keel_Length + 0.73053*(Femur_Length)^2**

| | Model 1 | Model 2 |
|---|---|---|
| Variables | Total Length, Wight, Humerus Length, Keel Length | Total Length, Wight, Humerus Length, Keel Length, Femur Length Square |
| 95 CI Interval | 2.5 % 97.5 %<br>(Intercept) 0.0891 1.5562<br>TotalL -3.9976 -1.3733<br>Wght -2.7848 -0.6810<br>HumerusL 1.1346 3.6088<br>KeelL 0.4649 2.4743 | 2.5 % 97.5 %<br>(Intercept) -0.9011 0.8460<br>TotalL -4.3168 -1.5369<br>Wght -2.4463 -0.5788<br>HumerusL 1.4181 4.1967<br>KeelL 0.4288 2.3287<br>I(FemurL^2) 0.1607 1.4699 |
| Outliers | 2 | 1 |
| 5-Fold CV Accuracy | 0.7882353 | 0.824183 |
| Proportion of Reduction in Error | 0.5335469 | 0.5669628 |
| AUC | 0.9251 | 0.93 |
| 95% CI AUC | [0.8733,0.977] | [0.878, 0.982] |

## Conclusion and Discussion

Since the all the coefficient intervals do not contain zero at 95% confidence level, the model does suggest total length, weight, Humerus length, keel length and square of fumur length have effect on the status of house sparrow. Further interpret for each coefficient, we have:

- When total length increase by 1 unit, the odd for house sparrow to survived is decreased by 93% ($e^{2.7}$)
- When weight increase by 1 unit, the odd for house sparrow to survive is decreased by 86% ($e^{-1.42}$).
- When humerus length increase by 1 unit, the odd for house sparrow to survived is increase by 12.36 times. ($e^{2.59}$)
- When Keel length increase by 1 unit, the odd for house sparrow to survived is increased by 2.6 times ($e^{1.28}$).
- When square of Femur lengths increase by 1 unit, the odd for house sparrow to survived is increased by 1.1 times ($e^{0.73}$).

A side note to the interpretation above, since we took natural log of weight and also scaled all variables, instead of using variable unit, we use unit-less unit in the interpretation.

We notice that some scientists also find out the width of skull may influence sparrows' survival status, but our final model does not show this aspect. To include this variable, we may need to do further research to improve the final model. Some people also use other method to fit the data, such as Structural equation modelling, so the logistic regression is not the only choice.

Appendix (all the plot)

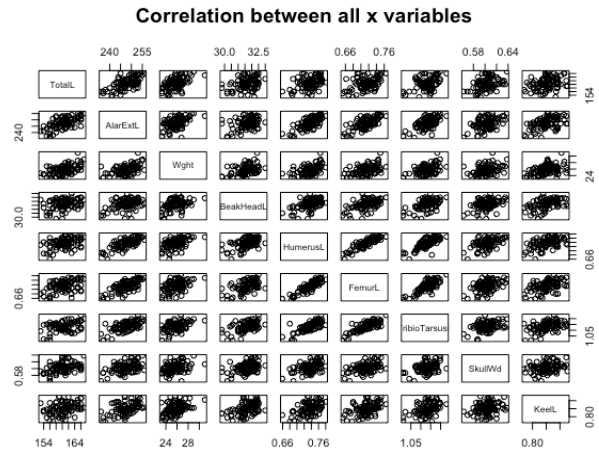


Figure 1

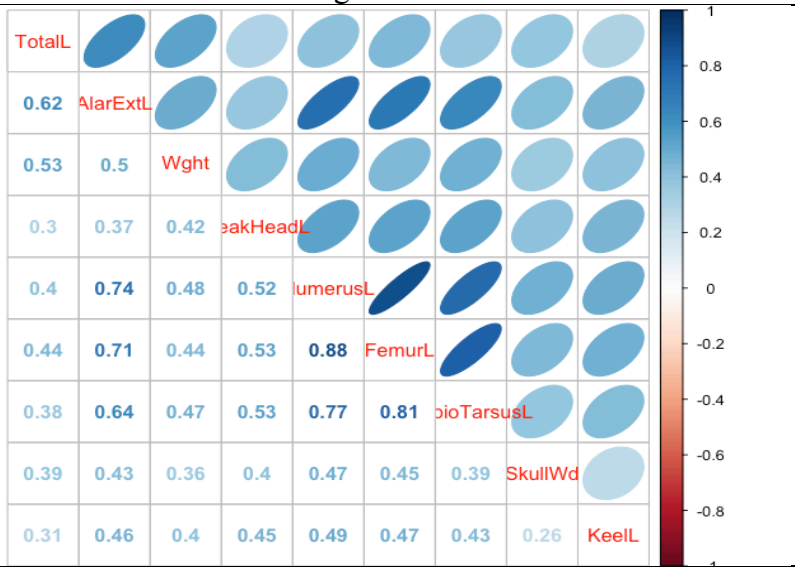


Figure 2

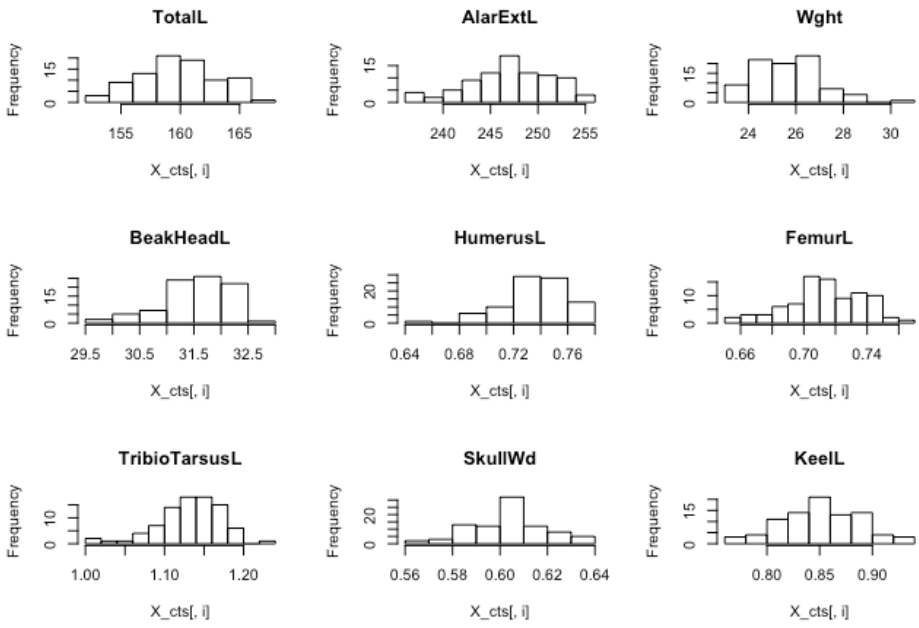


Figure 3

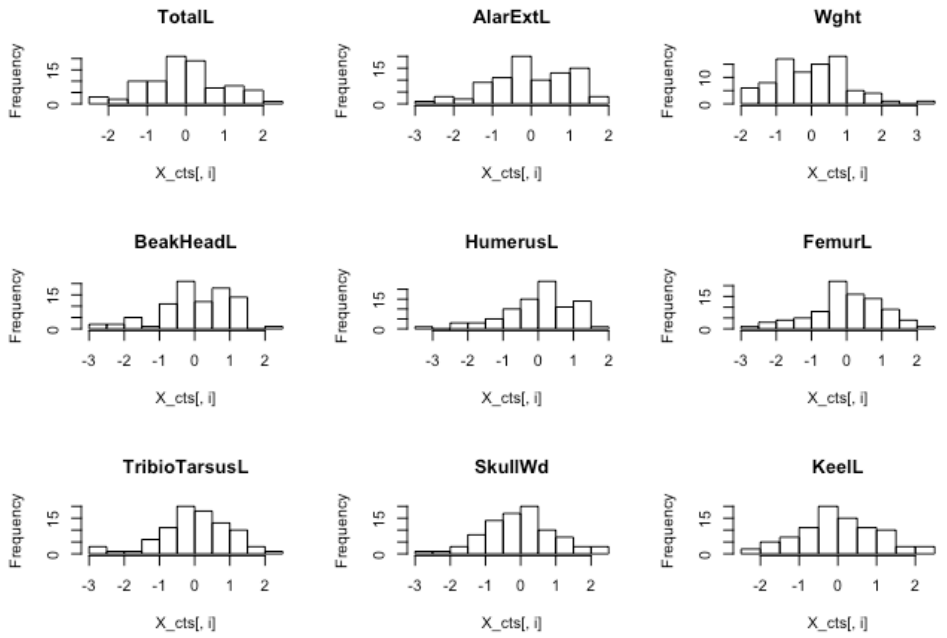


Figure 4

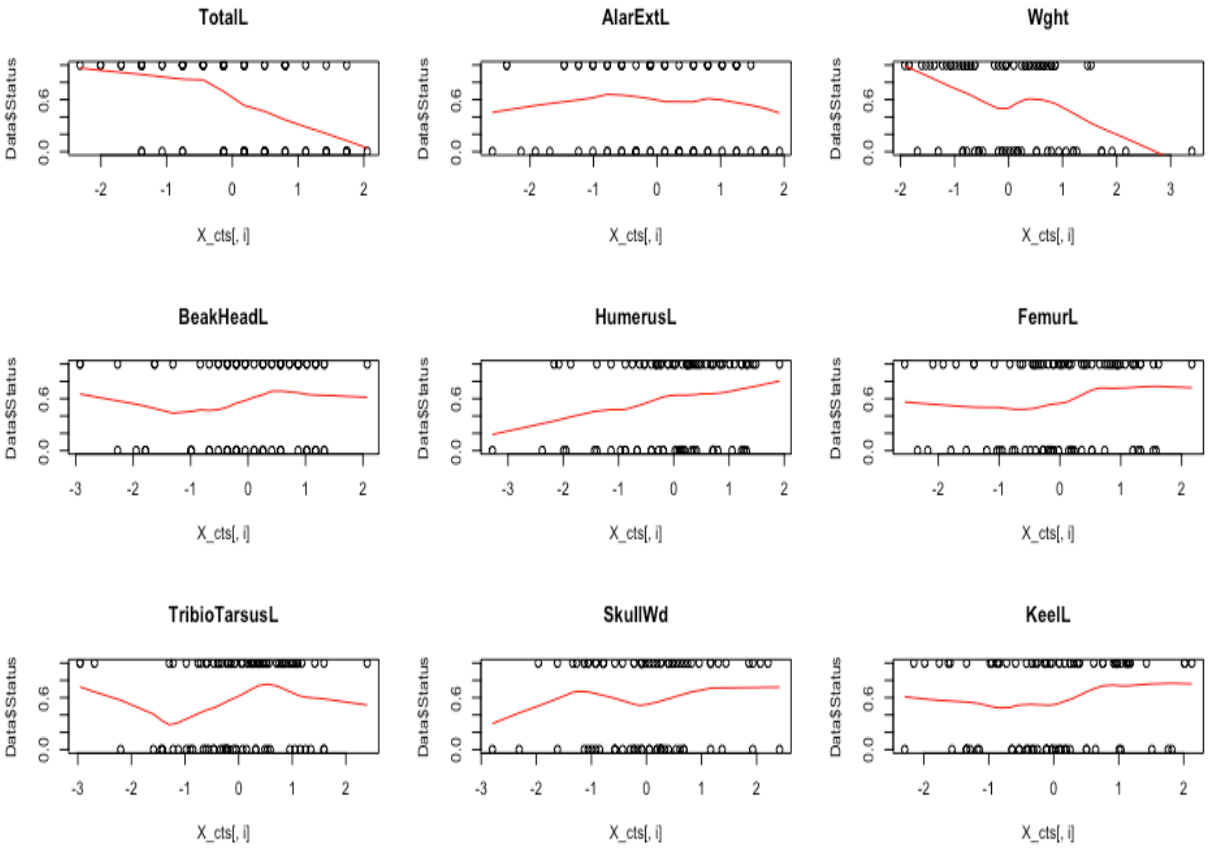


Figure 5

```
Call:  glm(formula = Status ~ TotalL + Wght + HumerusL + KeelL, family = binomial(link = logit),
    data = Data, maxit = 100)

Coefficients:
(Intercept)       TotalL         Wght     HumerusL        KeelL
     0.6335      -2.0292      -1.0690       1.6133       0.9245

Degrees of Freedom: 86 Total (i.e. Null);  82 Residual
Null Deviance:      118
Residual Deviance: 69.73        AIC: 79.73
```

Figure 6 (Best model for first-order)

```
Call:
glm(formula = Status ~ TotalL + Wght + HumerusL + KeelL, family = binomial(link = logit),
    data = Data, maxit = 100)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1992  -0.5898   0.2012   0.5855   2.2747

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.6335     0.3175   1.995 0.045998 *
TotalL       -2.0292     0.5231  -3.880 0.000105 ***
Wght         -1.0690     0.4192  -2.550 0.010776 *
HumerusL      1.6133     0.4709   3.426 0.000613 ***
KeelL         0.9245     0.4119   2.244 0.024820 *
---
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:  69.728  on 82  degrees of freedom
AIC: 79.728

Number of Fisher Scoring iterations: 6
```

Figure 7 (ANOVA table for best first-order model) (Reduced model)

```
Call:
glm(formula = Status ~ ., family = binomial(link = logit), data = Data,
    maxit = 100)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3233  -0.5891   0.1633   0.5501   1.8413

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.54096    0.95259   0.568 0.570116
Age            0.13324    0.68825   0.194 0.846491
TotalL        -2.43156    0.63435  -3.833 0.000126 ***
AlarExtL       0.45044    0.57482   0.784 0.433263
Wght          -1.20852    0.47067  -2.568 0.010238 *
BeakHeadL      0.36256    0.38958   0.931 0.352039
HumerusL       0.72456    0.79115   0.916 0.359754
FemurL         0.63145    0.85641   0.737 0.460928
TribioTarsusL -0.08847    0.58653  -0.151 0.880100
SkullWd        0.41900    0.39000   1.074 0.282654
KeelL          0.80872    0.43026   1.880 0.060163 .
---
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:  65.698  on 76  degrees of freedom
AIC: 87.698

Number of Fisher Scoring iterations: 6
```

Figure 8 (ANOVA table for full model in first-order) (Saturated model)

```
Call:  glm(formula = Status ~ TotalL + HumerusL + Wght + KeelL + I(FemurL^2),
    family = binomial(link = logit), data = Data)

Coefficients:
(Intercept)       TotalL     HumerusL         Wght        KeelL  I(FemurL^2)
     0.1472      -2.2760       2.0814      -1.1443       0.9354       0.5426

Degrees of Freedom: 86 Total (i.e. Null);  81 Residual
Null Deviance:      118
Residual Deviance: 65.22        AIC: 77.22
```

Figure 9 (Best model for second-order)

```
Call:
glm(formula = Status ~ TotalL + HumerusL + Wght + KeelL + I(FemurL^2),
    family = binomial(link = logit), data = Data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0920  -0.6015   0.1339   0.5021   2.6347

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.1472     0.3990   0.369 0.712218
TotalL       -2.2760     0.5786  -3.934 8.37e-05 ***
HumerusL      2.0814     0.5658   3.679 0.000234 ***
Wght         -1.1443     0.4186  -2.733 0.006268 **
KeelL         0.9354     0.4158   2.250 0.024456 *
I(FemurL^2)   0.5426     0.2826   1.920 0.054827 .
---
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:  65.223  on 81  degrees of freedom
AIC: 77.223

Number of Fisher Scoring iterations: 6
```

Figure 10 (ANOVA table for best second-order model)(Reduced model)

```
Call:
glm(formula = Status ~ . + I(TotalL^2) + I(AlarExtL^2) + I(Wght^2) +
    I(BeakHeadL^2) + I(HumerusL^2) + I(FemurL^2) + I(TribioTarsusL^2) +
    I(SkullWd^2) + I(KeelL^2), family = binomial(link = logit),
    data = Data, maxit = 100)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.57774  -0.38948   0.05348   0.40754   2.17289

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.1987     1.2415  -0.160  0.87281
Age                 0.1414     0.9769   0.145  0.88488
TotalL             -3.6903     1.0499  -3.515  0.00044 ***
AlarExtL            0.4928     0.6860   0.718  0.47250
Wght               -0.7485     0.5996  -1.248  0.21195
BeakHeadL           0.8477     0.5134   1.651  0.09868 .
HumerusL            0.6445     1.1019   0.585  0.55863
FemurL              1.0671     1.1358   0.939  0.34748
TribioTarsusL       0.2126     0.7643   0.278  0.78086
SkullWd             0.2439     0.4510   0.541  0.58861
KeelL               1.1451     0.6107   1.875  0.06079 .
I(TotalL^2)        -0.5933     0.6264  -0.947  0.34358
I(AlarExtL^2)      -0.4152     0.5106  -0.813  0.41611
I(Wght^2)          -0.2674     0.2589  -1.033  0.30180
I(BeakHeadL^2)      0.3686     0.2620   1.407  0.15940
I(HumerusL^2)      -0.8215     0.5784  -1.420  0.15554
I(FemurL^2)         1.9363     0.9984   1.939  0.05245 .
I(TribioTarsusL^2)  0.1076     0.5485   0.196  0.84455
I(SkullWd^2)       -0.1022     0.3429  -0.298  0.76560
I(KeelL^2)          0.7733     0.4654   1.662  0.09656 .
---
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:  51.814  on 67  degrees of freedom
AIC: 91.814
```

Figure 11 (ANOVA table for full model in second-order) (Saturated model)

Model 1 Residul Plot

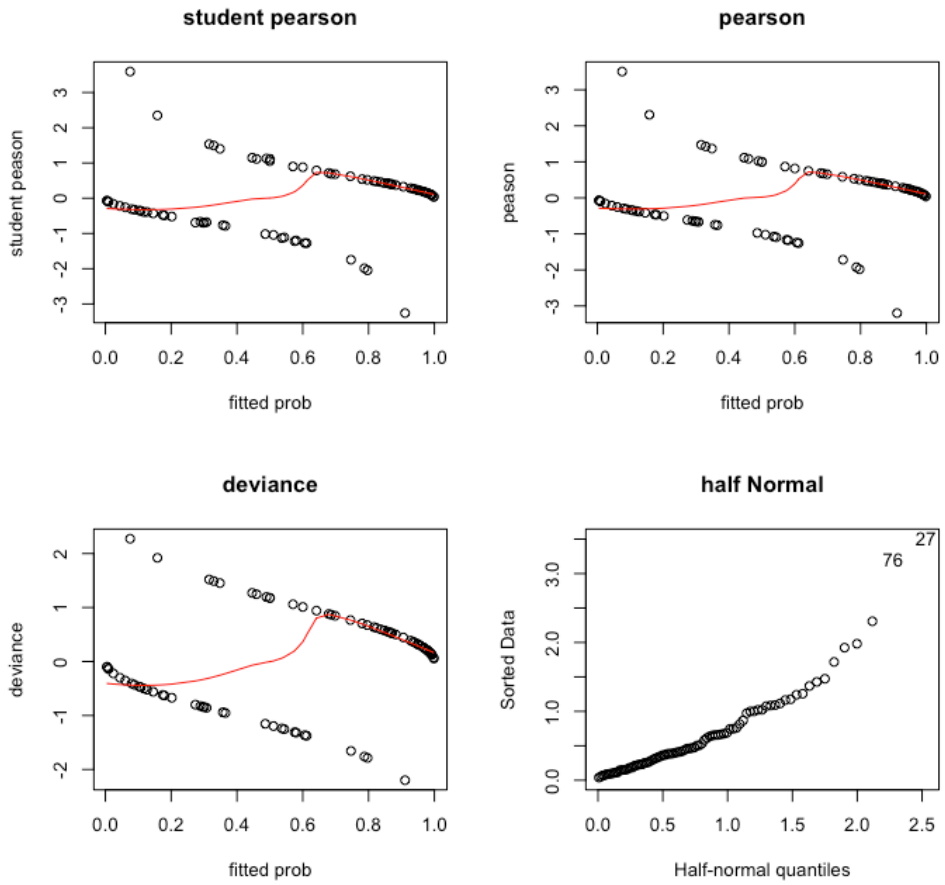


Figure 12

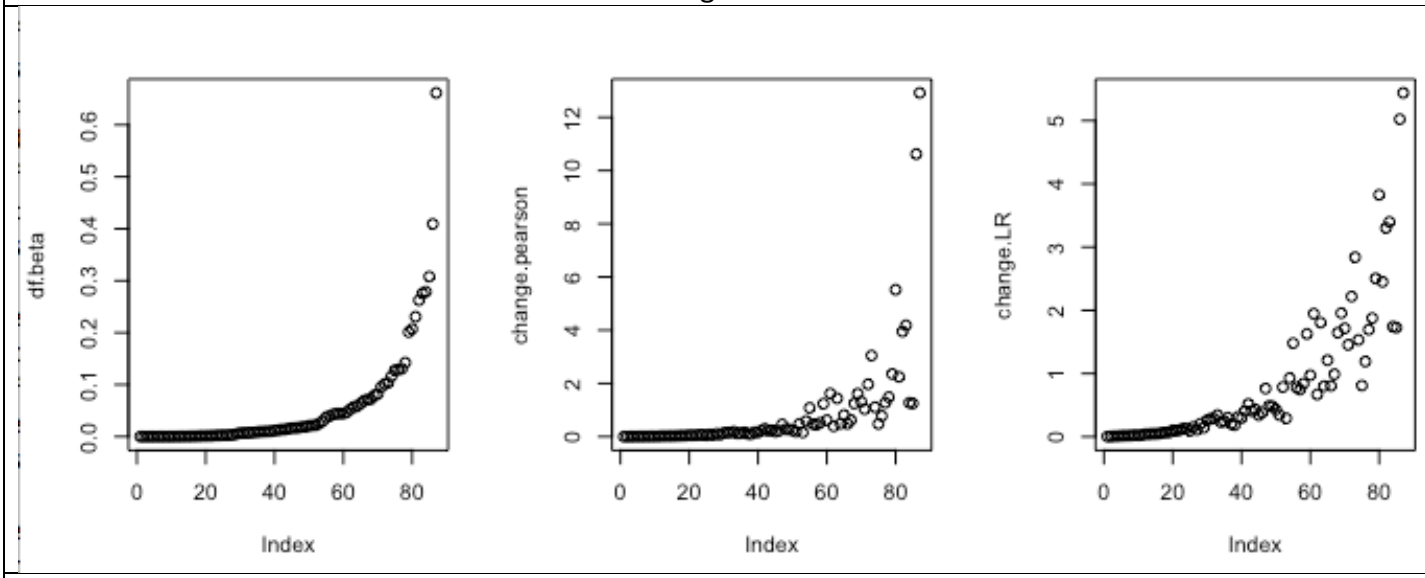


Figure 13

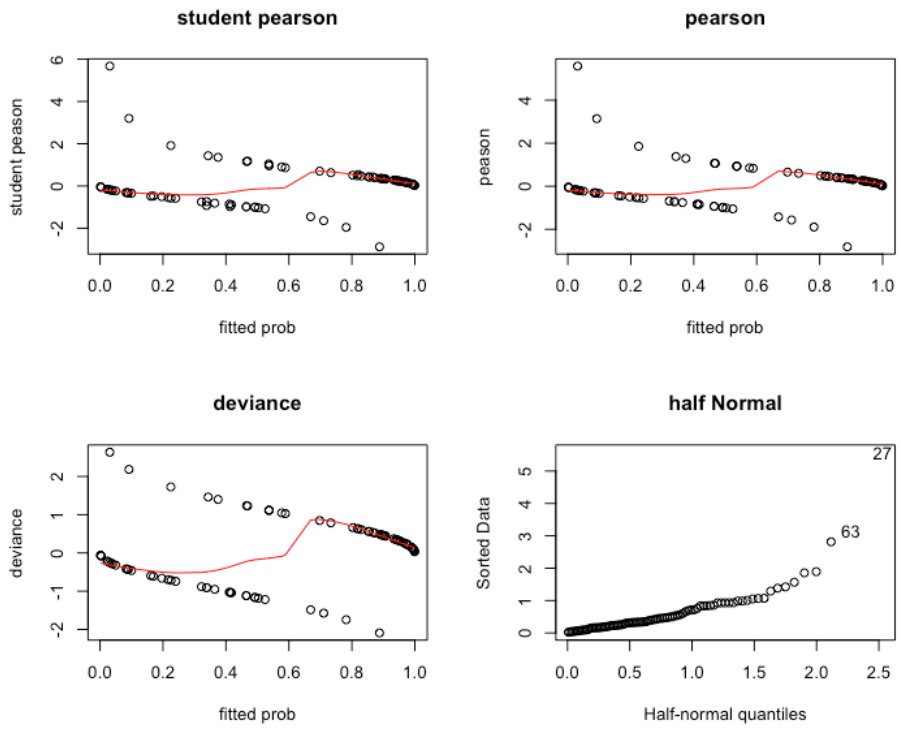


Figure 14

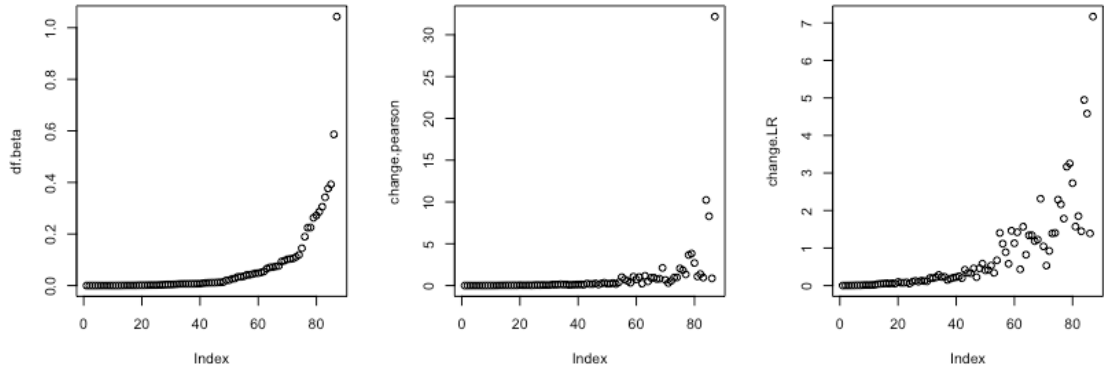


Figure 15

Appendix (all the R-code):

```
setwd("~/Desktop")
library(openxlsx)
library(corrplot)
Data = read.xlsx("survival_sparrow.xlsx")
names(Data)
names(Data)=c('Status','Age','TotalL','AlarExtL',
          'Wght','BeakHeadL','HumerusL',
          'FemurL','TribioTarsusL','SkullWd','KeelL')

# Data processing
#multicollinearity
dim(Data)
Data[Data$Status=='Perished',]$Status= 0
Data[Data$Status=='Survived',]$Status = 1
Data$Status = as.numeric(Data$Status)
Data$Age  = as.numeric(Data$Age )
sapply(Data,class)
X_cts = as.data.frame(Data[,3:11])
pairs(X_cts, main='Correlation between all x variables')
corrplot.mixed(cor(X_cts),lower = "number", upper="ellipse")

par(mfrow=c(3, 3))
for (i in 1:9){
  hist(X_cts[,i], main = names(Data)[i+2])
}

#check if we need to do transformation on X continuous variable
Data$Wght = log(X_cts$Wght)
Data$BeakHeadL= log(X_cts$BeakHeadL)


#standarized quantitative variable variables
for (i in 3:11){
  Data[,i] = (Data[,i]-mean(Data[,i]))/sd(Data[,i])
}
X_cts = as.data.frame(Data[,3:11])
par(mfrow=c(3, 3))
for (i in 1:9){
  hist(X_cts[,i], main = names(Data)[i+2])
}


par(mfrow=c(3, 3))
for (i in 1:9){
  plot(X_cts[,i],Data$Status, main = names(Data)[i+2] )
  lines(lowess(data.frame(X_cts[,i],Data$Status)), col = 2)
}



par(mfrow=c(1, 1))
```

```
# Fit with all fisrt-order
logit.model =  glm(formula = Status ~., family = binomial(logit), data = Data)
names(summary(logit.model))
summary(logit.model)
fitted(logit.model)
alpha = 0.05
confint(logit.model,level = (1-alpha))

#model slection stepwise/selection
glm.control(epsilon = 1e-8, maxit = 100, trace = FALSE)
full.model = glm(Status ~. , data =Data,family = binomial(link=logit),maxit = 100)
#View(data.frame(fitted(full.model ),Data$Status))
empty.model = glm(Status~ 1 ,data = Data,family = binomial(link=logit))

#Forward selection and backward selection with AIC
best.forward.AIC = step(empty.model,scope = list(lower = empty.model, upper = full.model),direction =
"forward", criterion = "AIC", trace = FALSE)
best.backward.AIC = step(full.model,scope = list(lower = empty.model, upper = full.model),direction =
"backward",  criterion = "AIC", trace = FALSE)
best.forward.AIC$formula # TotalL + Wght + HumerusL + KeelL
best.backward.AIC$formula # TotalL + Wght + HumerusL + KeelL

#Forward selection and backward selection with BIC
best.forward.BIC = step(empty.model,scope = list(lower = empty.model, upper = full.model),direction =
"forward", criterion = "AIC", k =log(87) , trace = FALSE)
best.backward.BIC = step(full.model,scope = list(lower = empty.model, upper = full.model),direction =
"backward",  criterion = "AIC",k =log(87) , trace = FALSE)
best.forward.BIC$formula #TotalL + Wght + HumerusL + KeelL
best.backward.BIC$formula #TotalL + Wght + HumerusL + KeelL

#summary(best.backward.AIC)
best.backward.AIC$formula
summary(full.model)

# Fit the model with second order
# remove the mean
for (i in 3:11){
  Data[,i] = Data[,i]-mean(Data[,i])
}

full.model = glm(Status ~. + I(TotalL^2)+I(AlarExtL^2)+ I(Wght^2)+ I(BeakHeadL^2)
            +I(HumerusL^2)+I(FemurL^2)+I(TribioTarsusL^2)+I(SkullWd^2)+I(KeelL^2) ,
               data =Data,family = binomial(link=logit),maxit = 100)
summary(full.model)
empty.model = glm(Status~ 1 ,data = Data,family = binomial(link=logit))
best.forward.AIC = step(empty.model,scope = list(lower = empty.model, upper = full.model),direction =
"forward", criterion = "AIC", trace = FALSE)
best.forward.BIC = step(empty.model,scope = list(lower = empty.model, upper = full.model),direction =
"forward", criterion = "AIC",k =log(87) ,trace = FALSE)
best.forward.AIC$formula
best.forward.AIC$formula
```

```
fullnew = best.forward.AIC
best.backward.AIC = step(fullnew,scope = list(lower = empty.model, upper = fullnew),direction =
"backward",  criterion = "AIC", trace = FALSE)
best.backward.BIC = step(fullnew,scope = list(lower = empty.model, upper = fullnew),direction =
"backward",  criterion = "AIC",k =log(87) , trace = FALSE)
best.backward.AIC$formula
best.backward.BIC$formula


# fit the model with interaction term
glm.control(epsilon = 1e-8, maxit = 100, trace = FALSE)
full.model = glm(Status ~. + .*. ,data =Data,family = binomial(link=logit),maxit = 100)
#View(data.frame(fitted(full.model ),Data$Status))
empty.model = glm(Status~ 1 ,data = Data,family = binomial(link=logit))
#Forward selection AIC
best.forward.AIC = step(empty.model,scope = list(lower = empty.model, upper = full.model),direction =
"forward", criterion = "AIC", trace = FALSE)
best.forward.AIC$formula #Status ~ TotalL + HumerusL + Wght + KeelL


#Forward selection BIC
best.forward.BIC = step(empty.model,scope = list(lower = empty.model, upper = full.model),direction =
"forward", criterion = "AIC", k =log(87) , trace = FALSE)
best.forward.BIC$formula #Status ~ TotalL + HumerusL + Wght + KeelL


setwd("~/Dropbox/0STA207/207 HW/207 Prj")
library(openxlsx)
library(corrplot)
Data = read.xlsx("survival_sparrow.xlsx")
names(Data)=c('Status','Age','TotalL','AlarExtL',
            'Wght','BeakHeadL','HumerusL',
            'FemurL','TribioTarsusL','SkullWd','KeelL')

Data[Data$Status=='Perished',]$Status= 0
Data[Data$Status=='Survived',]$Status = 1
Data$Status = factor(Data$Status)
Data$Age  = factor(Data$Age )
Data$Wght = log(Data$Wght)
Data$BeakHeadL= log(Data$BeakHeadL)
#standarized quantitative variable variables
for (i in 3:11){
  Data[,i] = (Data[,i]-mean(Data[,i]))/sd(Data[,i])
}

###4
#########################
#residual plot
logit.model = glm(Status ~ TotalL + Wght + HumerusL + KeelL, data =Data,family =
binomial(link=logit),maxit = 100)
```

```
Resi_Plot(logit.model )

library(LogisticDx)
par(mfrow=c(1, 3))
good.stuff = dx(logit.model)
df.beta = good.stuff$dBhat #DF Beta for removing each observation
plot(df.beta)
cutoff.beta = 0.5
df.beta[df.beta > cutoff.beta]
good.stuff[df.beta > cutoff.beta]
change.pearson = good.stuff$dChisq #Change in pearson X^2 for each observation
plot(change.pearson)
cutoff.pearson = 8
change.pearson[change.pearson > cutoff.pearson] #Shows the values
good.stuff[change.pearson > cutoff.pearson,] #what observations they were
#
change.LR = good.stuff$dDev #Change in LR-test G^2 for each observation
plot(change.LR)
good.stuff[change.LR> 4,] #what observations they were

Data = Data[-27,]
dim(Data)
Data = Data[-75,]
dim(Data)

best.model = glm(Status ~ TotalL + Wght + HumerusL + KeelL, data =Data,family =
binomial(link=logit),maxit = 100)
##cross validation
library(caret)
ctrl <- trainControl(method = "repeatedcv", number = 5, savePredictions = TRUE)
mod_fit <- train(Status ~ TotalL + Wght + HumerusL + KeelL ,  data=Data[,-c(4,9)], method="glm",
family=binomial(logit),
                 trControl = ctrl, tuneLength = 5)
mod_fit
#r^2
r = cor(best.model$y,best.model$fitted.values)
r
prop.red = 1- sum((best.model$y -best.model$fitted.values)^2)/sum((best.model$y -
mean(best.model$y))^2)
prop.red
#Classification tables, AUC, ROC.
library(pROC)
the.roc = roc(best.model$y, best.model$fitted.values,auc = TRUE, ci = TRUE,plot=TRUE, legacy.axes
= TRUE)
auc(the.roc)
ci(the.roc)
pi0 =0.50
my.table = table(truth = best.model$y,predict = ifelse(fitted(best.model)>pi0,1,0))
my.table
```

```
#and the AUC, plot the ROC, and find a confidence interval for the AUC. It requires the actual values of YY, the fitted values, and some arguments so that the AUC, and a confidence interval for the AUC are given back.

alpha = 0.05
round(confint(best.model,level = (1-alpha)),4)

###5
#########################
Data = read.xlsx("survival_sparrow.xlsx")
names(Data)=c('Status','Age','TotalL','AlarExtL',
              'Wght','BeakHeadL','HumerusL',
              'FemurL','TribioTarsusL','SkullWd','KeelL')

# Data processing
Data[Data$Status=='Perished',]$Status= 0
Data[Data$Status=='Survived',]$Status = 1
Data$Status = factor(Data$Status)
Data$Age  = factor(Data$Age )
Data$Wght = log(Data$Wght)
Data$BeakHeadL= log(Data$BeakHeadL)
#standarized quantitative variable variables
for (i in 3:11){
  Data[,i] = (Data[,i]-mean(Data[,i]))/sd(Data[,i])
}
logit.model = glm(Status ~ TotalL + Wght + HumerusL + KeelL + I(FemurL^2), data =Data,family =
binomial(link=logit),maxit = 100)
Resi_Plot(logit.model )

library(LogisticDx)
par(mfrow=c(1, 3))
good.stuff = dx(logit.model)
df.beta = good.stuff$dBhat #DF Beta for removing each observation
plot(df.beta)
cutoff.beta = 0.9
df.beta[df.beta > cutoff.beta]
good.stuff[df.beta > cutoff.beta]
change.pearson = good.stuff$dChisq #Change in pearson X^2 for each observation
plot(change.pearson)
cutoff.pearson = 15
change.pearson[change.pearson > cutoff.pearson] #Shows the values
good.stuff[change.pearson > cutoff.pearson,] #what observations they were
#
change.LR = good.stuff$dDev #Change in LR-test G^2 for each observation
plot(change.LR)
good.stuff[change.LR> 6,] #what observations they were

#27 63
Data[27,]
Data = Data[-27,]
dim(Data)
```

```
best.model = glm(Status ~ TotalL + Wght + HumerusL + KeelL + I(FemurL^2), data =Data,family =
binomial(link=logit),maxit = 100)
##cross validation
library(caret)
ctrl <- trainControl(method = "repeatedcv", number = 5, savePredictions = TRUE)
mod_fit <- train(Status ~TotalL + Wght + HumerusL + KeelL + I(FemurL^2) ,  data=Data[,-c(4,9)],
method="glm", family=binomial(logit),
                 trControl = ctrl, tuneLength = 5)
mod_fit
#r^2
r = cor(best.model$y,best.model$fitted.values)
r
prop.red = 1- sum((best.model$y -best.model$fitted.values)^2)/sum((best.model$y -
mean(best.model$y))^2)
prop.red
#Classification tables, AUC, ROC.
library(pROC)
the.roc = roc(best.model$y, best.model$fitted.values,auc = TRUE, ci = TRUE,plot=TRUE, legacy.axes
= TRUE)
auc(the.roc)
ci(the.roc)
pi0 =0.50
my.table = table(truth = best.model$y,predict = ifelse(fitted(best.model)>pi0,1,0))
my.table
#and the AUC, plot the ROC, and find a confidence interval for the AUC. It requires the actual values of
YY, the fitted values, and some arguments so that the AUC, and a confidence interval for the AUC are
given back.
Resi_Plot(best.model)
alpha = 0.05
round(confint(best.model,level = (1-alpha)),4)

exp(summary(best.model)$coefficients[,1])

Resi_Plot = function(logit.model){
  par(mfrow=c(2, 2))
  plot(fitted(logit.model),
       resid(logit.model, type='pearson')/sqrt(1 - hatvalues(logit.model)),
       main = 'student pearson', ylab = 'student peason',xlab = 'fitted prob')
  lines(lowess(data.frame(fitted(logit.model),
                          resid(logit.model, type='pearson')/sqrt(1 - hatvalues(logit.model)))), col = 2)
  plot(fitted(logit.model),
       resid(logit.model, type='pearson'),
       main = 'pearson', ylab = 'peason',xlab = 'fitted prob')
  lines(lowess(data.frame(fitted(logit.model),
                          resid(logit.model, type='pearson'))), col = 2)
  plot(fitted(logit.model),resid(logit.model),
       main = 'deviance', ylab = 'deviance',xlab = 'fitted prob')
  lines(lowess(data.frame(fitted(logit.model),
                          resid(logit.model, type='deviance'))), col = 2)
  library(faraway)
  faraway::halfnorm( resid(logit.model,'pearson') ,main = 'half Normal')
}
```