House sparrow survival data analysis

Abstract

this study is to analyze various characteristics of house sparrows who were found on the ground after a severe winter storm in 1898 and investigate if the probability of survival is associated with physical characteristics. Thus, we built a logistic regression model to find the relationship between survival status and these physical characteristics. Our final model shows that the probability of survival has nonlinear relationship with Total length (TL), Length of humerus (HL), Weight (WT) and Length of keel of sternum (KL). And the interpretation of odds shows that as the longer of the total length (TL) and the higher of the weight(WT), the sparrow is less likely to survive, but the sparrow is more likely to survive with the longer humerus(HL) and the longer keel of sternum.

1. Introduction

1.1 Data Description

the data we used to analyze includes various characteristics of house sparrows who were found on the ground after a severe winter storm in 1898. In this House sparrow survival data, response variable (Y) is survival status (survived = 1 or perished = 0), and predictor variables includes 1 qualitative variable : Age (adult = 1, juvenile = 2) and 9 quantitative variables : Total length (TL), Alar extent (AE), Weight (WT), Length of beak and head (BH), Length of humerus (HL), Lengths of femur (FL), Length of tibio-tarsus (TT), Width of skull (SK), Length of keel of sternum (KL). There are 87 observations in total.

1.2 Goal of study

The ecologists want to investigate if the probability of survival is associated with physical characteristics. Thus, the goal of our study is to build a logistic regression model to find the relationship between survival status and these physical characteristics.

2. Preliminary study and project plan

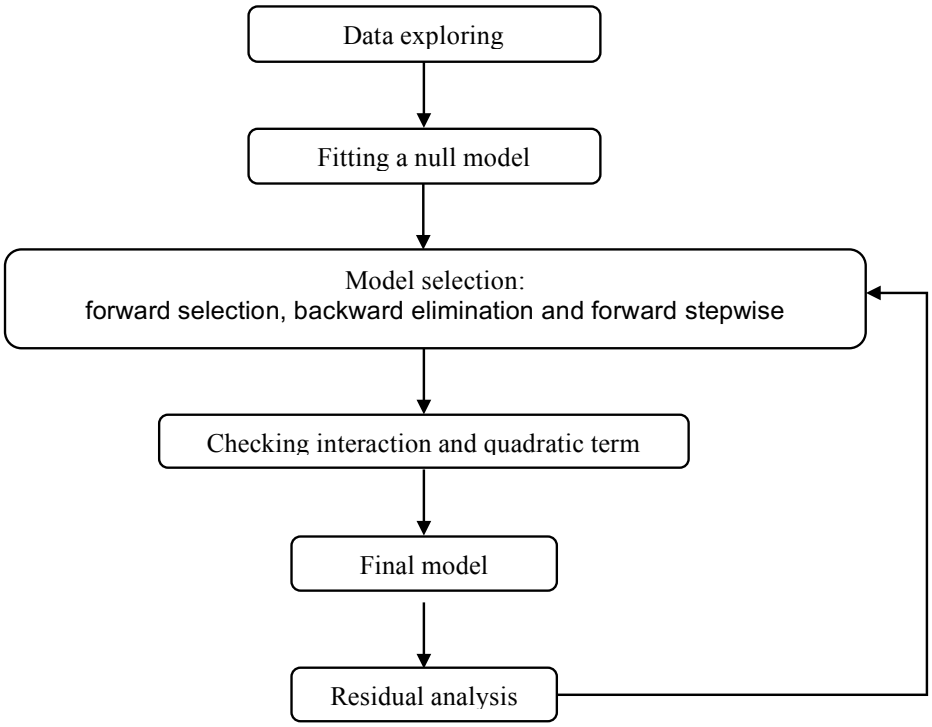
2.1 Data exploring

In the first step, we plot a scatterplot for each predictor in the dataset for detecting the existence of multicollinearity. According to (figure.1), we could detect the existence of multicollinearity such as AE&BH, FL&HL and FL&TT, etc. We will remove the multicollinearity by stepwise procedure in the following step.

2.2 Model and model selection

In this study, the response variable is qualitative with two possible outcomes: survival status (survived = 1 or perished = 0) . Therefore, we consider the logistic nonlinear regression model. We will use forward selection, backward elimination and forward stepwise to do model selection with AIC as criterion.

2.3 Project plan



3. Model selection and diagnostic

3.1 Model selection

Stepwise procedure is an efficient way to select the optimal model. We conduct forward selection, backward elimination and forward stepwise and we obtain the same model based on finding the minimum AIC. The minimum AIC is 78.612 and the model includes only four predicators: TL, HL, WT and KL. TL, HL are significant under the significance level of $\alpha = 0.001$, and WT and KL are significant under the significance level of $\alpha = 0.005$. In this case, we could select TL, HL, WT and KL as our main effects. Based on the results of selecting model with only four variables, it will remove the multicollinearity in some circumstances and avoid of the overfitting problem.

3.2 Fit two-way interactions and quadratic term

After fitting the model with the main effects, we should check whether the model should include any interaction terms and quadratic terms. So, in this step, we enlarge the scope of the variables that include two-way interaction terms and quadratic terms and then do the forward stepwise again to select model. We could conclude that the model does not have any improvements by adding two-way interaction terms and quadratic terms. So, the model only with main effect fits better.

3.3 Final model

The final model is the model selected by AIC stepwise procedure:

$$logit(\pi) = 49.9861 - 0.6573 \times TL + 72.3327 \times HL - 0.7896 \times WT + 27.3775 \times KL$$

Where π represents the probability of the survive rate for sparrows.

3.4Diagnostic

If a model is properly fitted, there should be no correlation between residuals and predictors and fitted values, we can see the horizontal straight line without curvature. The residual plots (figure. 2) shows that there's no trend between residuals and predictor and fitted values. Therefore, we could conclude that the model is a properly fit.

However, the residual plot can only reflect the overall model fit, in the following sections, we should find out outliers and influential observations. Based on the outlier test, we can see that there is no outlier as judged by Bonferonni p.

From the table of influential observations (Table.1) and the plot (figure.3) of influential observations. Influential observations may cause substantial changes in the estimated coefficients. Influential observations could be detected by high hat-values, long Cook's distance and large studendized residuals. When influential observation is dropped from the model, there will be a significant shift of the coefficient. For instance, observation 27 has the highest studentized residuals and the rather high Cook's distance but a moderate hat-value. If we drop observation 27, we can see that the estimated coefficients are changed. (Table.2)

Variable	*Est (Final model)*	*SE (Final model)*	*Est (Adjusted model)*	*SE (Adjusted model)*
Intercept	49.986	18.488	51.684	19.719
TL	-0.657	0.168	-0.744	0.191
HL	72.333	20.764	85.028	24.242
WT	-0.790	0.310	-0.955	0.347
KL	27.377	11.778	35.721	13.292

table.2

4. Conclusion

4.1Interpretations

The odds of the probability of survival when total length(TL) increases by 1 unit is $\exp(-0.657) = 0.518$ times of the odds of the probability of survival before. The 95% Wald confidence interval for the odds ratio is [0.373,0.720]. In other words, as the longer of the total length (TL), the sparrow is less likely to survive.

The odds of the probability of survival when length of humerus(HL) increases by 1 unit is $\exp(72.333) = 2.59\times10^{31}$ times of the odds of the probability of survival before. The 95% Wald confidence interval for the odds ratio is [5.48×10^{15},1.23×10^{49}]. In other words, as the longer of the length of humerus(HL), the sparrow is more likely to survive and the range of confidence interval is very wide, it provides with an evidence of strong association between HL and the probability of survival, which means the predictor has significant effects on predicting the probability of survival for sparrows.

The odds of the probability of survival when weight(WT) increases by 1 unit is $\exp(-0.790) = 0.454$ times of the odds of the probability of survival before. The 95% Wald confidence interval

for the odds ratio is [0.247,0.834]. In other words, as the higher of the weight(WT), the sparrow is less likely to survive.

The odds of the probability of survival when Length of keel of sternum (KL) increases by 1 unit is $\exp(27.377) = 7.76\times10^{11}$ times of the odds of the probability of survival before. The 95% Wald confidence interval for the odds ratio is [73.156,4.82×10^{9}]. In other words, as the longer the Length of keel of sternum (KL), the sparrow is more likely to survive and the range of confidence interval is very wide, it provides with an evidence of strong association between KL and the probability of survival, which means the predictor has significant effects on predicting the probability of survival for sparrows.

4.2 Summary

The goal of this study is to investigate if the probability of survival is associated with physical characteristics. We built the logistic nonlinear regression model to show the relationship between survival status and these physical characteristics in the house sparrow survival dataset. First, we fitted null model, and used forward selection, backward elimination and forward stepwise to do model selection. We got the same results that we could select TL, HL, WT and KL as our main effects which give us the minimum AIC. And then, we expanded our potential scope and fit our initial model with interaction and quadratic term within these four main effects. But the model does not have any improvements by adding two-way interaction terms and quadratic terms. Our final model is $logit(\pi) = 49.9861 - 0.6573\times TL + 72.3327\times HL - 0.7896\times WT + 27.3775\times KL$.

To check weather our model is fitted properly, we did Residuals Analysis to diagnostic our model. The residual plots show that there's no obvious trend between residuals and predictor or fitted values. Therefore, we could conclude that the model is a properly fit.

Next, we focus on observations of outlier, leverage and influence that may have significant impact on model building. Summary statistics for outlier, leverage and influence are studentized residuals, hat values and Cook's distance. According to Cook's distance, we considered observation 27 as influential observation. When influential observation 27 is dropped from the model, there was a significant shift of the coefficient. Thus, the adjusted model is: $logit(\pi) = 51.684 - 0.744\times TL + 85.028\times HL - 0.955\times WT + 35.721\times KL$.

Finally, we interpreted the odds of these four main effects, the results show that as the longer of the total length (TL) and the higher of the weight(WT), the sparrow is less likely to survive. But the sparrow is more likely to survive with the longer humerus(HL) and the longer keel of sternum.

4.3 Improvement

In our study, we basically achieved our goal to investigate if the probability of survival is associated with physical characteristics. Our final model showed the relationship between survival status and these physical characteristics in the house sparrow survival dataset. However, if we want to know the robustness of model, we need to collect more data to test our model, and also build advanced model to investigate.

Appendix

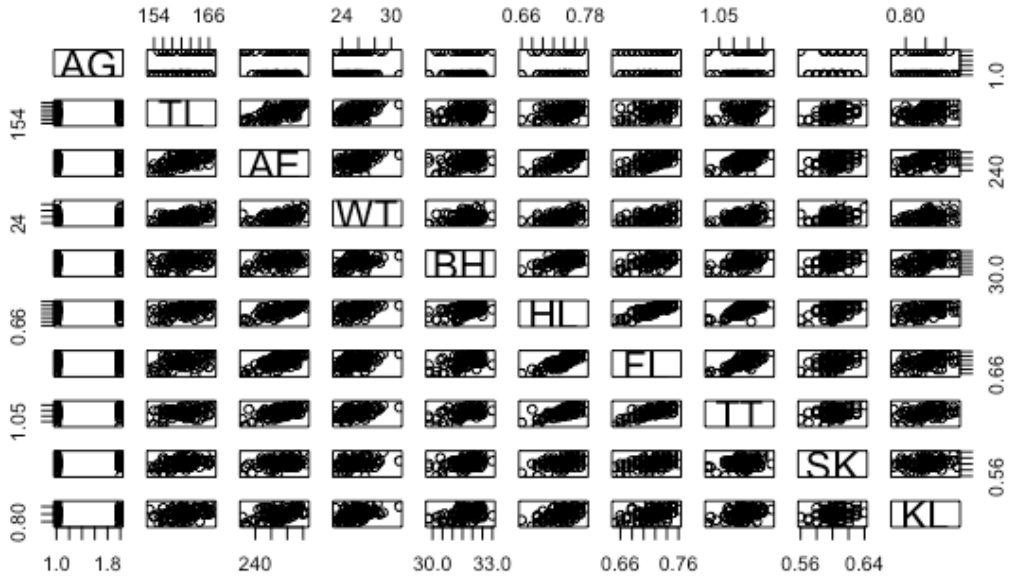


figure. 1

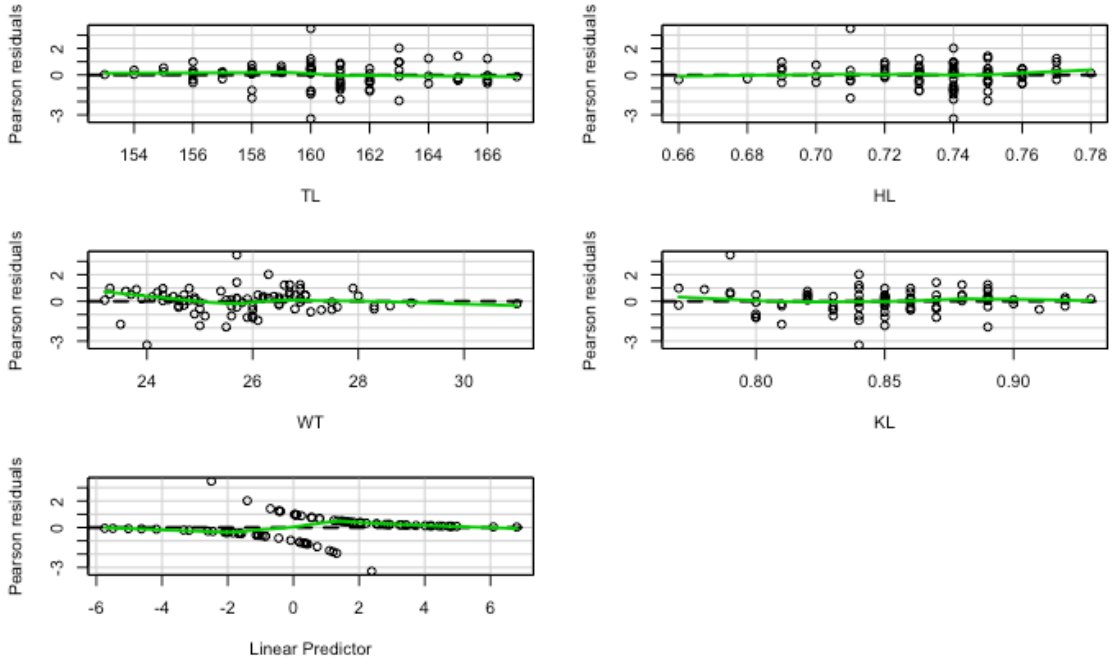


figure.2

	StudRes	*Hat*	*CookD*
27	2.4036074	0.04884504	0.13122582
32	1.1575261	0.18688699	0.04440425
40	-1.8384271	0.06668258	0.05727693
63	1.8473367	0.03884173	0.03425314
69	1.2231119	0.18368901	0.05004104
76	-2.3148894	0.03687071	0.08620248
77	-0.8189202	0.21097480	0.02282777

table 1

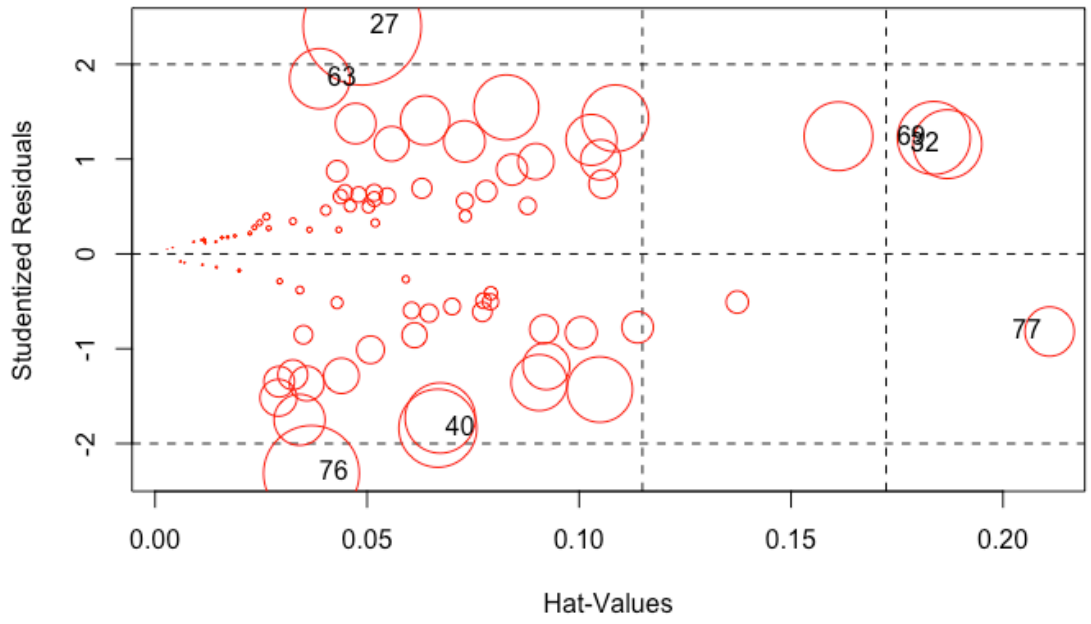


figure.3

Code

Dataset

```
library("gdata")
```

```
survival = read.xls("survival_sparrow.xls",header = T)
survival
```

```
##      STATUS AG  TL  AE   WT   BH   HL   FL   TT   SK   KL
## 1  Survived  1 154 241 24.5 31.2 0.69 0.67 1.02 0.59 0.83
## 2  Survived  1 160 252 26.9 30.8 0.74 0.71 1.18 0.60 0.84
## 3  Survived  1 155 243 26.9 30.6 0.73 0.70 1.15 0.60 0.85
## 4  Survived  1 154 245 24.3 31.7 0.74 0.69 1.15 0.58 0.84
## 5  Survived  1 156 247 24.1 31.5 0.71 0.71 1.13 0.57 0.82
## 6  Survived  1 161 253 26.5 31.8 0.78 0.74 1.14 0.61 0.89
## 7  Survived  1 157 251 24.6 31.1 0.74 0.74 1.15 0.61 0.86
## 8  Survived  1 159 247 24.2 31.4 0.73 0.72 1.13 0.61 0.79
## 9  Survived  1 158 247 23.6 29.8 0.70 0.67 1.08 0.60 0.82
## 10 Survived  1 158 252 26.2 32.0 0.75 0.74 1.15 0.61 0.86
## 11 Survived  1 160 252 26.2 32.0 0.74 0.72 1.13 0.62 0.89
## 12 Survived  1 162 253 24.8 32.3 0.77 0.75 1.13 0.63 0.92
## 13 Survived  1 161 243 25.4 31.8 0.72 0.72 1.13 0.60 0.89
## 14 Survived  1 160 250 23.7 29.8 0.73 0.70 1.10 0.59 0.82
## 15 Survived  1 159 247 25.7 31.4 0.73 0.72 1.14 0.59 0.93
## 16 Survived  1 158 253 25.7 31.9 0.74 0.70 1.15 0.60 0.86
## 17 Survived  1 159 247 26.5 31.6 0.73 0.71 1.15 0.61 0.92
## 18 Survived  1 166 253 26.7 32.5 0.77 0.76 1.23 0.60 0.88
## 19 Survived  1 159 247 23.9 31.4 0.75 0.72 1.11 0.60 0.82
## 20 Survived  1 160 248 24.7 31.3 0.75 0.74 1.18 0.60 0.80
## 21 Survived  1 161 252 28.0 31.8 0.77 0.73 1.19 0.59 0.88
## 22 Survived  1 163 251 27.9 31.9 0.77 0.75 1.17 0.62 0.86
## 23 Survived  1 156 242 25.9 32.0 0.72 0.71 1.12 0.61 0.89
## 24 Survived  1 165 251 25.7 32.2 0.75 0.74 1.16 0.61 0.87
## 25 Survived  1 160 247 26.6 32.4 0.73 0.71 1.11 0.59 0.84
## 26 Survived  1 158 244 23.2 31.6 0.73 0.71 1.14 0.58 0.89
## 27 Survived  1 160 242 25.7 31.6 0.71 0.70 1.12 0.62 0.79
## 28 Survived  1 157 245 26.3 32.2 0.74 0.73 1.14 0.60 0.85
## 29 Survived  1 159 244 24.3 31.5 0.72 0.70 1.11 0.62 0.85
## 30 Survived  1 160 253 26.7 32.1 0.74 0.71 1.12 0.59 0.86
## 31 Survived  1 158 245 24.9 31.4 0.73 0.70 1.12 0.58 0.85
## 32 Survived  1 161 247 23.8 31.4 0.74 0.69 1.10 0.60 0.78
## 33 Survived  1 160 247 25.6 32.3 0.76 0.75 1.13 0.61 0.90
## 34 Survived  1 160 247 27.0 32.0 0.75 0.74 1.17 0.63 0.87
## 35 Survived  1 153 241 24.7 32.2 0.73 0.68 1.09 0.59 0.88
## 36 Perished  1 165 249 26.5 31.0 0.74 0.70 1.10 0.61 0.85
## 37 Perished  1 160 245 26.1 32.0 0.74 0.71 1.11 0.61 0.84
## 38 Perished  1 161 249 25.6 32.3 0.74 0.72 1.13 0.60 0.83
## 39 Perished  1 162 246 25.9 32.3 0.74 0.71 1.13 0.61 0.87
## 40 Perished  1 163 250 25.5 32.5 0.75 0.73 1.20 0.62 0.89
## 41 Perished  1 162 247 27.6 31.8 0.73 0.72 1.11 0.60 0.87
## 42 Perished  1 163 246 25.8 31.4 0.69 0.66 1.07 0.60 0.84
## 43 Perished  1 161 246 24.9 30.5 0.74 0.73 1.14 0.58 0.80
```

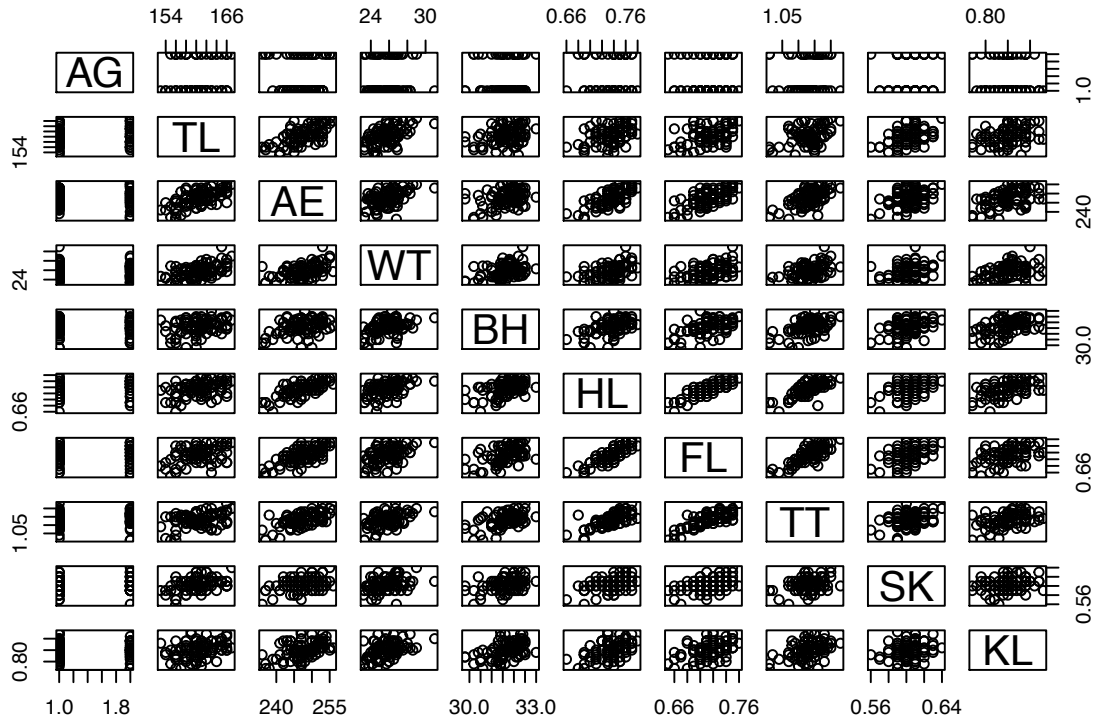
```
## 44 Perished  1 160 242 26.0 31.0 0.75 0.71 1.11 0.60 0.80
## 45 Perished  1 162 246 26.5 31.5 0.72 0.70 1.09 0.61 0.81
## 46 Perished  1 160 249 26.0 31.4 0.73 0.69 1.10 0.60 0.85
## 47 Perished  1 161 250 27.1 31.6 0.74 0.71 1.12 0.63 0.85
## 48 Perished  1 162 248 25.1 31.9 0.74 0.72 1.15 0.59 0.84
## 49 Perished  1 165 252 26.0 32.3 0.73 0.71 1.14 0.61 0.89
## 50 Perished  1 161 243 25.6 32.5 0.71 0.71 1.12 0.61 0.83
## 51 Perished  1 161 244 25.0 31.3 0.70 0.69 1.08 0.60 0.87
## 52 Perished  1 162 248 24.6 31.0 0.71 0.70 1.09 0.59 0.84
## 53 Perished  1 164 244 25.0 31.2 0.70 0.69 1.07 0.61 0.80
## 54 Perished  1 158 247 26.0 32.0 0.73 0.71 1.14 0.61 0.80
## 55 Perished  1 162 253 28.3 31.8 0.75 0.72 1.15 0.60 0.86
## 56 Perished  1 156 239 24.6 30.5 0.66 0.66 1.04 0.57 0.81
## 57 Perished  1 166 251 27.5 31.5 0.72 0.69 1.12 0.61 0.85
## 58 Perished  1 165 253 31.0 32.4 0.76 0.75 1.18 0.61 0.90
## 59 Perished  1 166 250 28.3 32.4 0.75 0.72 1.18 0.61 0.92
## 60 Survived  2 156 246 24.6 32.0 0.74 0.74 1.17 0.59 0.85
## 61 Survived  2 156 245 25.5 32.1 0.76 0.72 1.15 0.62 0.82
## 62 Survived  2 163 248 24.8 32.2 0.74 0.73 1.16 0.61 0.85
## 63 Survived  2 163 248 26.3 33.0 0.74 0.70 1.15 0.61 0.84
## 64 Survived  2 160 250 24.4 31.5 0.75 0.71 1.17 0.60 0.89
## 65 Survived  2 156 237 23.3 30.6 0.69 0.66 1.01 0.59 0.77
## 66 Survived  2 162 253 26.7 32.0 0.76 0.73 1.20 0.63 0.88
## 67 Survived  2 163 254 26.4 32.0 0.77 0.75 1.16 0.61 0.89
## 68 Survived  2 164 251 26.9 32.0 0.75 0.74 1.17 0.62 0.89
## 69 Survived  2 163 244 24.3 31.3 0.72 0.68 1.08 0.61 0.89
## 70 Survived  2 160 247 27.0 31.5 0.76 0.73 1.18 0.62 0.85
## 71 Survived  2 160 250 26.8 32.5 0.76 0.73 1.12 0.63 0.84
## 72 Survived  2 158 247 24.9 32.4 0.75 0.72 1.14 0.59 0.87
## 73 Survived  2 158 249 26.1 32.2 0.74 0.74 1.15 0.60 0.82
## 74 Survived  2 158 243 26.6 32.4 0.75 0.71 1.16 0.61 0.89
## 75 Survived  2 155 237 23.3 30.2 0.69 0.65 1.01 0.59 0.79
## 76 Perished  2 160 249 24.0 30.4 0.74 0.72 1.13 0.62 0.84
## 77 Perished  2 156 236 26.8 30.2 0.69 0.67 1.07 0.56 0.83
## 78 Perished  2 158 240 23.5 31.0 0.71 0.70 1.11 0.60 0.81
## 79 Perished  2 166 245 26.9 31.7 0.71 0.69 1.11 0.60 0.85
## 80 Perished  2 165 255 28.6 31.5 0.77 0.74 1.17 0.61 0.85
## 81 Perished  2 157 238 24.7 31.2 0.68 0.68 1.16 0.60 0.77
## 82 Perished  2 164 250 27.3 31.8 0.76 0.73 1.17 0.59 0.86
## 83 Perished  2 166 256 25.7 31.7 0.75 0.75 1.19 0.60 0.86
## 84 Perished  2 167 255 29.0 32.2 0.76 0.75 1.20 0.64 0.86
## 85 Perished  2 161 246 25.0 31.5 0.74 0.71 1.12 0.59 0.85
## 86 Perished  2 166 254 27.5 31.4 0.76 0.74 1.12 0.60 0.91
## 87 Perished  2 161 251 26.0 31.5 0.73 0.71 1.12 0.59 0.83
```

```
sapply(survival,class)
```

```
##    STATUS        AG        TL        AE        WT        BH        HL
##  "factor" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
##        FL        TT        SK        KL
## "numeric" "numeric" "numeric" "numeric"
```

```
survival$AG=as.factor(survival$AG)
```

```
plot(survival[,2:11])
```




```
cor(survival[,3:11])
```

```
##            TL        AE        WT        BH        HL        FL        TT
## TL 1.0000000 0.6216688 0.5251925 0.2971997 0.3873714 0.4188333 0.3714058
## AE 0.6216688 1.0000000 0.4990499 0.3740344 0.7340034 0.6961208 0.6271673
## WT 0.5251925 0.4990499 1.0000000 0.4208610 0.4675116 0.4440005 0.4689626
## BH 0.2971997 0.3740344 0.4208610 1.0000000 0.5189299 0.5335587 0.5193747
## HL 0.3873714 0.7340034 0.4675116 0.5189299 1.0000000 0.8440408 0.7373573
## FL 0.4188333 0.6961208 0.4440005 0.5335587 0.8440408 1.0000000 0.7909904
## TT 0.3714058 0.6271673 0.4689626 0.5193747 0.7373573 0.7909904 1.0000000
## SK 0.4051501 0.3947013 0.3480838 0.4028379 0.4371073 0.4258273 0.3547988
## KL 0.3225277 0.4623729 0.4081211 0.4638314 0.4860805 0.4555498 0.4193330
##            SK        KL
## TL 0.4051501 0.3225277
## AE 0.3947013 0.4623729
## WT 0.3480838 0.4081211
## BH 0.4028379 0.4638314
## HL 0.4371073 0.4860805
## FL 0.4258273 0.4555498
## TT 0.3547988 0.4193330
## SK 1.0000000 0.2452865
## KL 0.2452865 1.0000000
```

null model (forward)

```
survival$STATUS = as.factor(survival$STATUS)
fit.null=glm(STATUS~1, data=survival,
             family=binomial)
fit.forward=step(fit.null,
```

```
                scope=~AG+TL+AE+WT+BH+HL+FL+TT+SK+KL,
                direction='forward')
```

```
## Start:  AIC=120.01
## STATUS ~ 1
##
##        Df Deviance    AIC
## + TL    1   99.788 103.79
## + WT    1  111.249 115.25
## + HL    1  114.431 118.43
## <none>     118.008 120.01
## + KL    1  116.321 120.32
## + FL    1  116.521 120.52
## + TT    1  116.942 120.94
## + BH    1  117.090 121.09
## + SK    1  117.854 121.85
## + AG    1  117.971 121.97
## + AE    1  117.986 121.99
##
## Step:  AIC=103.79
## STATUS ~ TL
##
##        Df Deviance     AIC
## + HL    1   80.020  86.020
## + FL    1   85.223  91.223
## + AE    1   87.703  93.703
## + KL    1   89.162  95.162
## + TT    1   89.618  95.618
## + BH    1   92.820  98.820
## + SK    1   93.844  99.844
## <none>      99.788 103.788
## + WT    1   99.546 105.546
## + AG    1   99.755 105.755
##
## Step:  AIC=86.02
## STATUS ~ TL + HL
##
##        Df Deviance    AIC
## + WT    1   75.094 83.094
## + KL    1   76.708 84.708
## <none>      80.020 86.020
## + SK    1   78.537 86.537
## + BH    1   78.846 86.846
## + AE    1   79.656 87.656
## + TT    1   79.775 87.775
## + FL    1   79.857 87.857
## + AG    1   80.020 88.020
##
## Step:  AIC=83.09
## STATUS ~ TL + HL + WT
##
##        Df Deviance    AIC
## + KL    1   68.612 78.612
## + BH    1   72.512 82.512
```

```
## <none>      75.094 83.094
## + SK    1   73.451 83.451
## + AE    1   74.450 84.450
## + TT    1   74.460 84.460
## + FL    1   74.789 84.789
## + AG    1   75.076 85.076
##
## Step:  AIC=78.61
## STATUS ~ TL + HL + WT + KL
##
##        Df Deviance    AIC
## <none>      68.612 78.612
## + BH    1   67.214 79.214
## + SK    1   67.496 79.496
## + TT    1   68.206 80.206
## + AE    1   68.334 80.334
## + FL    1   68.541 80.541
## + AG    1   68.612 80.612
```

```
summary(fit.forward)
```

```
##
## Call:
## glm(formula = STATUS ~ TL + HL + WT + KL, family = binomial,
##     data = survival)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2234  -0.5648   0.1540   0.6094   2.2701
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  49.9861    18.4879   2.704 0.006857 **
## TL           -0.6573     0.1683  -3.907 9.35e-05 ***
## HL           72.3327    20.7640   3.484 0.000495 ***
## WT           -0.7896     0.3097  -2.549 0.010800 *
## KL           27.3775    11.7780   2.324 0.020101 *
## ---
## 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:  68.612  on 82  degrees of freedom
## AIC: 78.612
##
## Number of Fisher Scoring iterations: 6
```

Backward method

```
fit.full=glm(STATUS~., data=survival,
             family=binomial)
fit.back = step(fit.full, scope=~AG+TL+AE+WT+BH+HL+FL+TT+SK+KL,
                direction='back')
```

```
## Start:  AIC=87.92
```

```
## STATUS ~ AG + TL + AE + WT + BH + HL + FL + TT + SK + KL
##
##        Df Deviance     AIC
## - AG    1   65.945  85.945
## - FL    1   65.964  85.964
## - TT    1   66.048  86.048
## - AE    1   66.358  86.358
## - SK    1   66.553  86.553
## - BH    1   66.867  86.867
## <none>      65.920  87.920
## - HL    1   69.494  89.494
## - KL    1   70.308  90.308
## - WT    1   74.894  94.894
## - TL    1   90.731 110.731
##
## Step:  AIC=85.94
## STATUS ~ TL + AE + WT + BH + HL + FL + TT + SK + KL
##
##        Df Deviance     AIC
## - FL    1   65.992  83.992
## - TT    1   66.074  84.074
## - AE    1   66.360  84.360
## - SK    1   66.592  84.592
## - BH    1   66.879  84.879
## <none>      65.945  85.945
## - HL    1   69.596  87.596
## - KL    1   70.314  88.314
## - WT    1   74.986  92.986
## - TL    1   91.219 109.219
##
## Step:  AIC=83.99
## STATUS ~ TL + AE + WT + BH + HL + TT + SK + KL
##
##        Df Deviance     AIC
## - TT    1   66.075  82.075
## - AE    1   66.410  82.410
## - SK    1   66.661  82.661
## - BH    1   66.884  82.884
## <none>      65.992  83.992
## - KL    1   70.317  86.317
## - HL    1   70.433  86.433
## - WT    1   74.993  90.993
## - TL    1   91.567 107.567
##
## Step:  AIC=82.07
## STATUS ~ TL + AE + WT + BH + HL + SK + KL
##
##        Df Deviance     AIC
## - AE    1   66.581  80.581
## - SK    1   66.758  80.758
## - BH    1   67.145  81.145
## <none>      66.075  82.075
## - KL    1   70.417  84.417
## - HL    1   71.852  85.852
```

```
## - WT    1   74.998  88.998
## - TL    1   91.628 105.628
##
## Step:  AIC=80.58
## STATUS ~ TL + WT + BH + HL + SK + KL
##
##        Df Deviance     AIC
## - SK    1   67.214  79.214
## - BH    1   67.496  79.496
## <none>      66.581  80.581
## - KL    1   71.652  83.652
## - WT    1   75.161  87.161
## - HL    1   79.250  91.250
## - TL    1   94.091 106.091
##
## Step:  AIC=79.21
## STATUS ~ TL + WT + BH + HL + KL
##
##        Df Deviance     AIC
## - BH    1   68.612  78.612
## <none>      67.214  79.214
## - KL    1   72.512  82.512
## - WT    1   76.295  86.295
## - HL    1   82.170  92.170
## - TL    1   94.183 104.183
##
## Step:  AIC=78.61
## STATUS ~ TL + WT + HL + KL
##
##        Df Deviance     AIC
## <none>      68.612  78.612
## - KL    1   75.094  83.094
## - WT    1   76.708  84.708
## - HL    1   86.690  94.690
## - TL    1   94.701 102.701
```

```
summary(fit.back)
```

```
##
## Call:
## glm(formula = STATUS ~ TL + WT + HL + KL, family = binomial,
##     data = survival)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2234  -0.5648   0.1540   0.6094   2.2701
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  49.9861    18.4879   2.704 0.006857 **
## TL           -0.6573     0.1683  -3.907 9.35e-05 ***
## WT           -0.7896     0.3097  -2.549 0.010800 *
## HL           72.3327    20.7640   3.484 0.000495 ***
## KL           27.3775    11.7780   2.324 0.020101 *
## ---
```

```
## 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:  68.612  on 82  degrees of freedom
## AIC: 78.612
##
## Number of Fisher Scoring iterations: 6
```

Both method

```
fit.both = step(fit.null, scope=~AG+TL+AE+WT+BH+HL+FL+TT+SK+KL,
                direction='both')
```

```
## Start:  AIC=120.01
## STATUS ~ 1
##
##        Df Deviance    AIC
## + TL    1   99.788 103.79
## + WT    1  111.249 115.25
## + HL    1  114.431 118.43
## <none>     118.008 120.01
## + KL    1  116.321 120.32
## + FL    1  116.521 120.52
## + TT    1  116.942 120.94
## + BH    1  117.090 121.09
## + SK    1  117.854 121.85
## + AG    1  117.971 121.97
## + AE    1  117.986 121.99
##
## Step:  AIC=103.79
## STATUS ~ TL
##
##        Df Deviance     AIC
## + HL    1   80.020  86.020
## + FL    1   85.223  91.223
## + AE    1   87.703  93.703
## + KL    1   89.162  95.162
## + TT    1   89.618  95.618
## + BH    1   92.820  98.820
## + SK    1   93.844  99.844
## <none>      99.788 103.788
## + WT    1   99.546 105.546
## + AG    1   99.755 105.755
## - TL    1  118.008 120.008
##
## Step:  AIC=86.02
## STATUS ~ TL + HL
##
##        Df Deviance     AIC
## + WT    1   75.094  83.094
## + KL    1   76.708  84.708
## <none>      80.020  86.020
## + SK    1   78.537  86.537
```

```
## + BH    1   78.846  86.846
## + AE    1   79.656  87.656
## + TT    1   79.775  87.775
## + FL    1   79.857  87.857
## + AG    1   80.020  88.020
## - HL    1   99.788 103.788
## - TL    1  114.431 118.431
##
## Step:  AIC=83.09
## STATUS ~ TL + HL + WT
##
##        Df Deviance     AIC
## + KL    1   68.612  78.612
## + BH    1   72.512  82.512
## <none>      75.094  83.094
## + SK    1   73.451  83.451
## + AE    1   74.450  84.450
## + TT    1   74.460  84.460
## + FL    1   74.789  84.789
## + AG    1   75.076  85.076
## - WT    1   80.020  86.020
## - TL    1   97.273 103.273
## - HL    1   99.546 105.546
##
## Step:  AIC=78.61
## STATUS ~ TL + HL + WT + KL
##
##        Df Deviance     AIC
## <none>      68.612  78.612
## + BH    1   67.214  79.214
## + SK    1   67.496  79.496
## + TT    1   68.206  80.206
## + AE    1   68.334  80.334
## + FL    1   68.541  80.541
## + AG    1   68.612  80.612
## - KL    1   75.094  83.094
## - WT    1   76.708  84.708
## - HL    1   86.690  94.690
## - TL    1   94.701 102.701
```

```
summary(fit.both)
```

```
##
## Call:
## glm(formula = STATUS ~ TL + HL + WT + KL, family = binomial,
##     data = survival)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2234  -0.5648   0.1540   0.6094   2.2701
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  49.9861    18.4879   2.704 0.006857 **
## TL           -0.6573     0.1683  -3.907 9.35e-05 ***
```

```
## HL           72.3327    20.7640   3.484 0.000495 ***
## WT           -0.7896     0.3097  -2.549 0.010800 *
## KL           27.3775    11.7780   2.324 0.020101 *
## ---
## 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:  68.612  on 82  degrees of freedom
## AIC: 78.612
##
## Number of Fisher Scoring iterations: 6
```

Based on the R outputs, we can fit model with four main effects

```
fit.final = glm(STATUS~TL+HL+WT+KL, data = survival, family = binomial)
summary(fit.final)
```

```
##
## Call:
## glm(formula = STATUS ~ TL + HL + WT + KL, family = binomial,
##     data = survival)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2234  -0.5648   0.1540   0.6094   2.2701
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  49.9861    18.4879   2.704 0.006857 **
## TL           -0.6573     0.1683  -3.907 9.35e-05 ***
## HL           72.3327    20.7640   3.484 0.000495 ***
## WT           -0.7896     0.3097  -2.549 0.010800 *
## KL           27.3775    11.7780   2.324 0.020101 *
## ---
## 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:  68.612  on 82  degrees of freedom
## AIC: 78.612
##
## Number of Fisher Scoring iterations: 6
```

Two-way interaction terms and Quadratic terms

```
fit.2way=step(fit.final, scope=~(TL+HL+WT+KL)^2+I(TL^2)+I(HL^2)+I(WT^2)+I(KL^2),
              direction = 'both')
```

```
## Start:  AIC=78.61
## STATUS ~ TL + HL + WT + KL
```

```
##
##           Df Deviance     AIC
## <none>         68.612  78.612
## + I(KL^2)  1   66.841  78.841
## + HL:WT    1   66.944  78.944
## + HL:KL    1   67.184  79.184
## + I(TL^2)  1   67.441  79.441
## + I(HL^2)  1   67.825  79.825
## + WT:KL    1   67.856  79.856
## + TL:HL    1   67.870  79.870
## + TL:KL    1   68.321  80.321
## + TL:WT    1   68.428  80.428
## + I(WT^2)  1   68.487  80.487
## - KL       1   75.094  83.094
## - WT       1   76.708  84.708
## - HL       1   86.690  94.690
## - TL       1   94.701 102.701
```

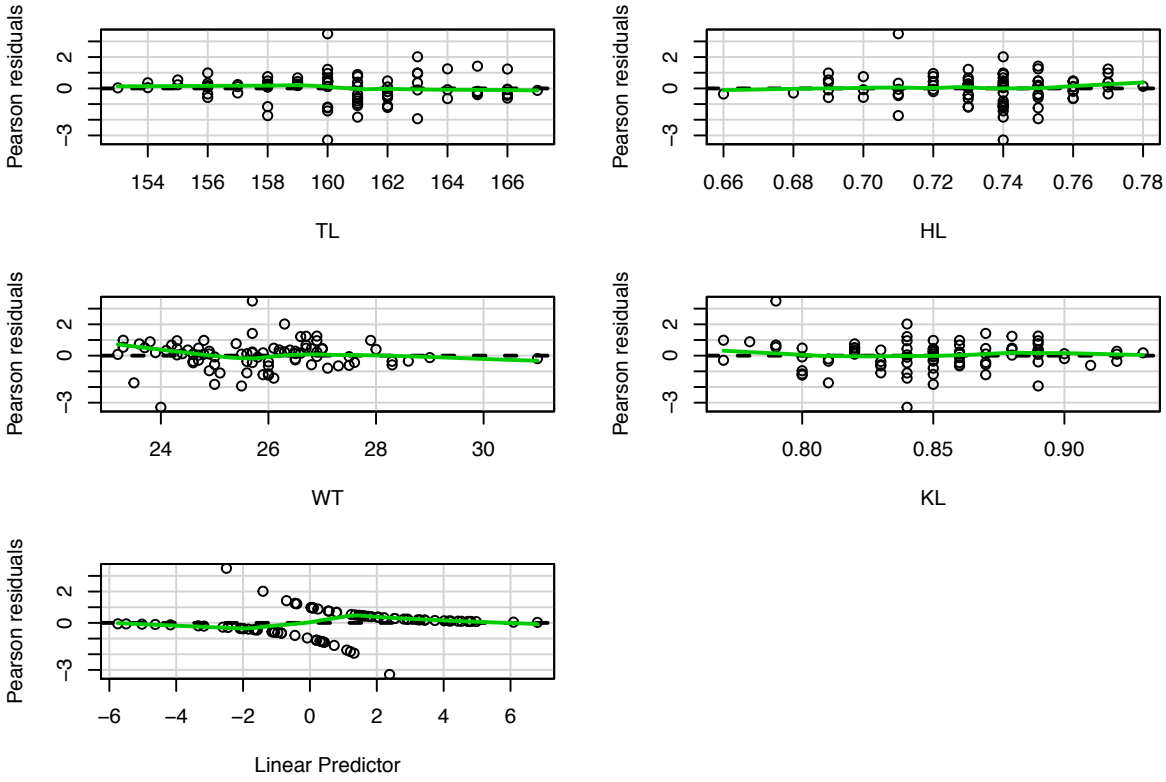
```
summary(fit.2way)
```

```
##
## Call:
## glm(formula = STATUS ~ TL + HL + WT + KL, family = binomial,
##     data = survival)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2234  -0.5648   0.1540   0.6094   2.2701
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  49.9861    18.4879   2.704 0.006857 **
## TL           -0.6573     0.1683  -3.907 9.35e-05 ***
## HL           72.3327    20.7640   3.484 0.000495 ***
## WT           -0.7896     0.3097  -2.549 0.010800 *
## KL           27.3775    11.7780   2.324 0.020101 *
## ---
## 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:  68.612  on 82  degrees of freedom
## AIC: 78.612
##
## Number of Fisher Scoring iterations: 6
```

Residual Analysis

```
library(car)
```

```
residualPlots(fit.final)
```




```
##    Test stat Pr(>|t|)
## TL     1.171    0.279
## HL     0.787    0.375
## WT     0.125    0.724
## KL     1.771    0.183
```

The residual plot can only reflect the overal model fit. The following sections tend to find out outliers, leverage and influence.

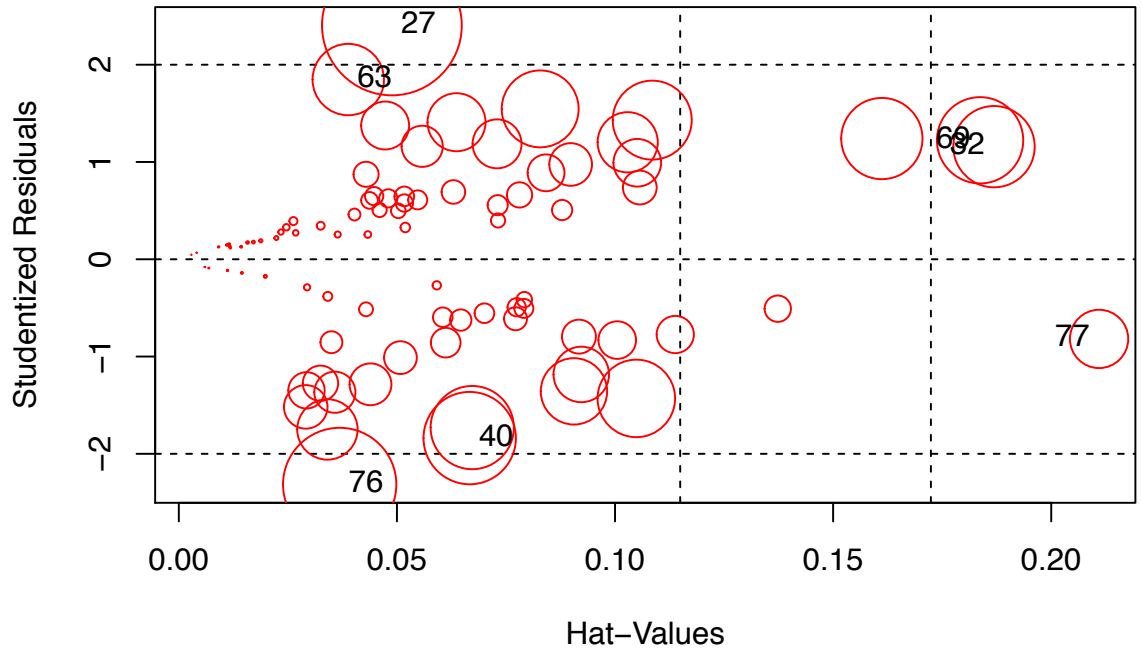
Outliers

```
outlierTest(fit.final)
```

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferonni p
## 27 2.403607           0.016234           NA
```

influence plot

```
influencePlot(fit.final, col = "red",id.n =3)
```




```
##       StudRes        Hat      CookD
## 27  2.4036074 0.04884504 0.13122582
## 32  1.1575261 0.18688699 0.04440425
## 40 -1.8384271 0.06668258 0.05727693
## 63  1.8473367 0.03884173 0.03425314
## 69  1.2231119 0.18368901 0.05004104
## 76 -2.3148894 0.03687071 0.08620248
## 77 -0.8189202 0.21097480 0.02282777
```

```
model69 = update(fit.final,subset=c(-27))
compareCoefs(fit.final,model69)
```

```
##
## Call:
## 1: glm(formula = STATUS ~ TL + HL + WT + KL, family = binomial, data =
##   survival)
## 2: glm(formula = STATUS ~ TL + HL + WT + KL, family = binomial, data =
##   survival, subset = c(-27))
##             Est. 1   SE 1 Est. 2   SE 2
## (Intercept) 49.986 18.488 51.684 19.719
## TL          -0.657  0.168 -0.744  0.191
## HL          72.333 20.764 85.028 24.242
## WT          -0.790  0.310 -0.955  0.347
## KL          27.377 11.778 35.721 13.292
```

Interpretations

```
#TL
exp(fit.final$coef[2])
```

```
##       TL
## 0.518223
```

```r
# lower bound
exp(fit.final$coef[2]-1.96*0.168)
```

```
##        TL
## 0.3728311
```

```r
# upper bound
exp(fit.final$coef[2]+1.96*0.168)
```

```
##        TL
## 0.7203128
```

```r
#HL
exp(fit.final$coef[3])
```

```
##           HL
## 2.592281e+31
```

```r
# lower bound
exp(fit.final$coef[3]-1.96*20.764)
```

```
##           HL
## 5.482877e+13
```

```r
# upper bound
exp(fit.final$coef[3]+1.96*20.764)
```

```
##          HL
## 1.22562e+49
```

```r
#WT
exp(fit.final$coef[4])
```

```
##        WT
## 0.4540375
```

```r
# lower bound
exp(fit.final$coef[4]-1.96*0.310)
```

```
##        WT
## 0.2472944
```

```r
# upper bound
exp(fit.final$coef[4]+1.96*0.310)
```

```
##        WT
## 0.8336217
```

```r
#KL
exp(fit.final$coef[5])
```

```
##           KL
## 776040490414
```

```r
# lower bound
exp(fit.final$coef[5]-1.96*11.778)
```

```
##       KL
## 73.15574
```

```r
# upper bound
exp(fit.final$coef[4]+1.96*11.778)*1000
```

```
##           WT
## 4.816457e+12
```