STAT 226 Lecture 15

Case Study: Bumpus Nature Selection Data

Yibi Huang

Example: Bumpus Nature Selection Data

In 1899, biologist Hermon Bumpus presented as evidence of natural selection a comparison of numerical characteristics of 87 moribund house sparrows that were collected after an uncommonly severe winter storm and which had either perished or survived as a result of their injuries.

Bumpus asked whether some sparrows were more likely to die because they lacked some physical characteristics that enables them to withstand the intensity of the storm[1].

```
bumpus = read.table(
  "http://www.stat.uchicago.edu/~yibi/s226/Bumpus.txt",
  h = TRUE)
```

[1] Data from Exercise 16 in Chapter 20 of *The Statistical Sleuth*, 3ed, by Ramsey and Schafer

Bumpus Data Variables

- `Status` Survival status, factor with levels "Perished" and "Survived"
- `AG`: age, factor with 2 levels: "adult" and "juvenile"
- `TL`: total length (in mm)
- `WT`: weight (in grams)
- `BH`: length of beak and head (in mm)
- `HL`: length of humerus (arm bone) (in inches)
- `FL`: length of femur (in inches)
- `TT`: length of tibio–tarsus (in inches)
- `SK`: width of skull (in inches)
- `KL`: length of keel of sternum (in inches)

Logistic Regression Using TL as the Only Predictor

`Status` cannot be used directly as the response.

```
glm(Status ~ TL, family=binomial, data=bumpus)
Error in eval(family$initialize): y values must be 0 <= y <= 1
```

Logistic Regression Using `TL` as the Only Predictor

`Status` cannot be used directly as the response.

```
glm(Status ~ TL, family=binomial, data=bumpus)
Error in eval(family$initialize): y values must be 0 <= y <= 1
```

Need to convert the levels of `Status` to 0 and 1, or to specify the "Success" category.

```
bumpus$Survived = as.numeric(bumpus$Status=="Survived")
glm(Survived ~ TL, family=binomial, data=bumpus)$coef
(Intercept)          TL
     54.493      -0.337
glm((Status == "Survived") ~ TL, family=binomial, data=bumpus)$coef
(Intercept)          TL
     54.493      -0.337
```

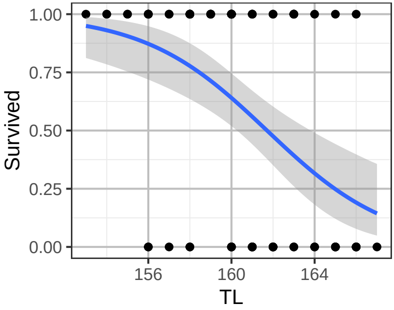
Fitted model: $\widehat{\pi}(x) = \dfrac{e^{54.493-0.337x}}{1 + e^{54.493-0.337x}}$.

```r
bumpus.fit1 = glm(Survived ~ TL, family=binomial, data=bumpus)
summary(bumpus.fit1)

Call:
glm(formula = Survived ~ TL, family = binomial, data = bumpus)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-2.030  -1.068   0.522   0.944   1.820

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  54.4931    14.5787    3.74  0.00019
TL           -0.3370     0.0906   -3.72  0.00020

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 118.008  on 86  degrees of freedom
Residual deviance:  99.788  on 85  degrees of freedom
AIC: 103.8
```

The function `geom_smooth()` in the `ggplot()` can overlay the fitted logistic curve on the scatter plot.

```
library(ggplot2)
ggplot(bumpus, aes(x=TL, y = Survived)) + geom_point() +
  geom_smooth(method = "glm", method.args = list(family = "binomial"))
```

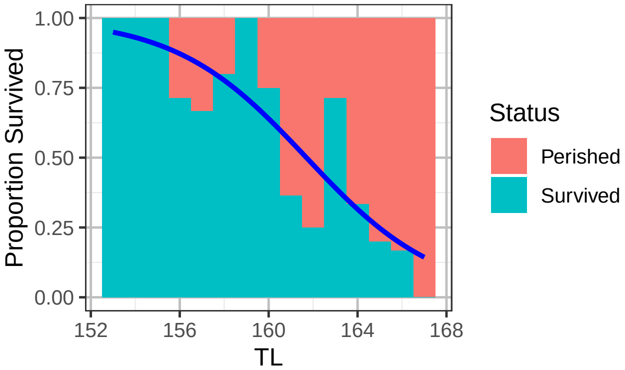
Hard to visually gauge how well the curve fits the data using such a plot.



```r
ggplot(bumpus, aes(x=TL, fill = Status)) +
  geom_histogram(binwidth=1) + theme(legend.position="top")
```

```r
xtabs(~ TL + Status, data=bumpus)
     Status
TL    Perished Survived
  153        0        1
  154        0        2
  155        0        2
  156        2        5
  157        1        2
  158        2        8
  159        0        5
  160        4       12
  161        7        4
  162        6        2
  163        2        5
  164        2        1
  165        4        1
  166        5        1
  167        1        0
```

```r
ggplot(bumpus, aes(x=TL, fill = Status))+
  geom_histogram(position = "fill", binwidth=1) +
  ylab("Proportion Survived") +
  geom_function(
      fun = function(x){exp(54.493-0.337*x)/(1+exp(54.493-0.337*x))},
      lwd=1,color="blue"
  )
```



```
prop.table(xtabs(~ TL + Status, data=bumpus),1)
     Status
TL     Perished Survived
  153    0.0000   1.0000
  154    0.0000   1.0000
  155    0.0000   1.0000
  156    0.2857   0.7143
  157    0.3333   0.6667
  158    0.2000   0.8000
  159    0.0000   1.0000
  160    0.2500   0.7500
  161    0.6364   0.3636
  162    0.7500   0.2500
  163    0.2857   0.7143
  164    0.6667   0.3333
  165    0.8000   0.2000
  166    0.8333   0.1667
  167    1.0000   0.0000
```

Fitted Logistic Regression Model

$$\widehat{\pi}(x) = \frac{\exp(\widehat{\alpha} + \widehat{\beta}x)}{1 + \exp(\widehat{\alpha} + \widehat{\beta}x)} = \frac{\exp(54.493 - 0.337x)}{1 + \exp(54.493 - 0.337x)}$$

- $\widehat{\beta} = -0.337 < 0$, so $\widehat{\pi}$ decreases as Total Length ($x = \texttt{TL}$) increases $\Rightarrow$ Longer birds are less likely to survive

Fitted Logistic Regression Model

$$\widehat{\pi}(x) = \frac{\exp(\widehat{\alpha} + \widehat{\beta}x)}{1 + \exp(\widehat{\alpha} + \widehat{\beta}x)} = \frac{\exp(54.493 - 0.337x)}{1 + \exp(54.493 - 0.337x)}$$

- $\widehat{\beta} = -0.337 < 0$, so $\widehat{\pi}$ decreases as Total Length ($x = \texttt{TL}$) increases $\Rightarrow$ Longer birds are less likely to survive
- Odds of survival were $e^{-0.337} \approx 0.714$ times as large for birds 1 mm longer in total length (`TL`)

Fitted Logistic Regression Model

$$\widehat{\pi}(x) = \frac{\exp(\widehat{\alpha} + \widehat{\beta}x)}{1 + \exp(\widehat{\alpha} + \widehat{\beta}x)} = \frac{\exp(54.493 - 0.337x)}{1 + \exp(54.493 - 0.337x)}$$

- $\widehat{\beta} = -0.337 < 0$, so $\widehat{\pi}$ decreases as Total Length ($x =$ `TL`) increases $\Rightarrow$ Longer birds are less likely to survive
- Odds of survival were $e^{-0.337} \approx 0.714$ times as large for birds 1 mm longer in total length (`TL`)
- Point of symmetry:

$$\widehat{\pi}(x) = \frac{1}{2} \text{ when } x = -\frac{\widehat{\alpha}}{\widehat{\beta}} = -\frac{54.493}{-0.337} = 161.7\,\text{mm}$$

Safer to Use the Likelihood Ratio CIs

95% Likelihood Ratio CI for β:

```
confint(bumpus.fit1, "TL", level=0.95)
Waiting for profiling to be done...
 2.5 % 97.5 %
-0.531 -0.172
```

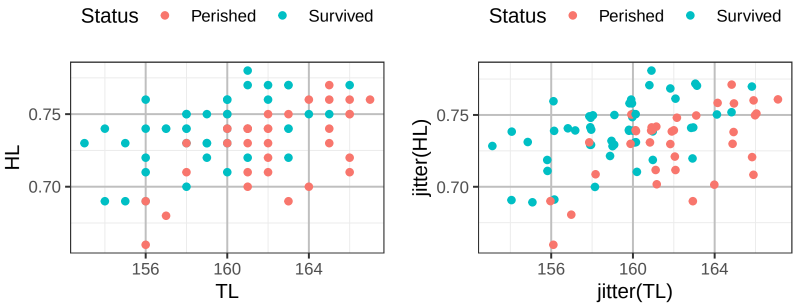
95% Likelihood Ratio CI for e^{β}:

```
exp(confint(bumpus.fit1, "TL", level=0.95))
Waiting for profiling to be done...
 2.5 % 97.5 %
 0.588  0.842
```

Interpretation: With 95% confidence, odds of survival become $e^{-0.531} \approx 0.588$ to $e^{-0.172} \approx 0.843$ times as large when the bird was 1 mm longer in total length (TL)

Taking Humerus Length (HL) Into Account

```r
ggplot(bumpus, aes(x=TL, y = HL, color=Status)) +
  geom_point()+theme(legend.position="top")
ggplot(bumpus, aes(x=jitter(TL), y = jitter(HL), color=Status)) +
  geom_point()+theme(legend.position="top")
```



Consider sparrows with the same TL (total length) were those with longer Humerus (arm bone) more likely to survive?

Model with Both HL and TL as Predictors

```
bumpus.fit2 = glm(Survived ~ TL + HL, family=binomial, data=bumpus)
summary(bumpus.fit2)$coef
             Estimate Std. Error z value    Pr(>|z|)
(Intercept)  54.0427    16.3906   3.297 0.000976648
TL           -0.6167     0.1393  -4.427 0.000009563
HL           61.7429    16.6296   3.713 0.000204957
```

Fitted Model:

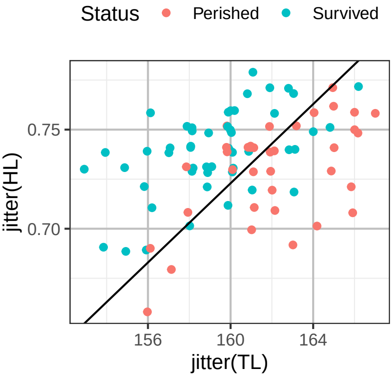
$$\widehat{\pi}(x) = \frac{\exp(54.043 - 0.6167\text{TL} + 61.743\text{HL})}{1 + \exp(54.043 - 0.6167\text{TL} + 61.743\text{HL})}$$

- odds of survival become $e^{61.743 \times 0.01} \approx 1.85$ times as large if the humerus is 0.01 inches longer for birds with the same total length
- odds of survival become $e^{-0.6167} \approx 0.54$ times as large if the bird is 1 mm longer in total length for birds with the same humerus length

The values of `TL` and `HL` that satisfies

$$54.043 - 0.617\text{TL} + 61.743\text{HL} = 0$$

are those with $\widehat{\pi} = 0.5$.



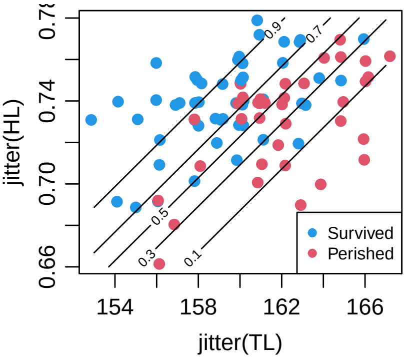
Points to the right/left of the line have less/greater than 50% estimated probability of survival.

Level Curves of Estimated Probabilities

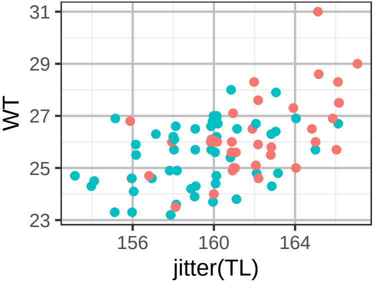
As $\widehat{\pi} = \frac{\exp(54.043-0.6167\mathrm{TL}+61.743\mathrm{HL})}{1+\exp(54.043-0.6167\mathrm{TL}+61.743\mathrm{HL})}$, observe

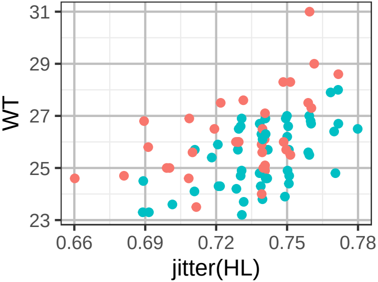
$$\begin{aligned}\widehat{\pi} = c \iff &\exp(54.043 - 0.6167\mathrm{TL} + 61.743\mathrm{HL}) = \frac{c}{1-c}\\ \iff &54.043 - 0.6167\mathrm{TL} + 61.743\mathrm{HL} = \log\left(\frac{c}{1-c}\right)\end{aligned}$$

The ($\mathtt{TL}$, $\mathtt{HL}$) values with $\widehat{\pi} = c$ are those on the straight line above



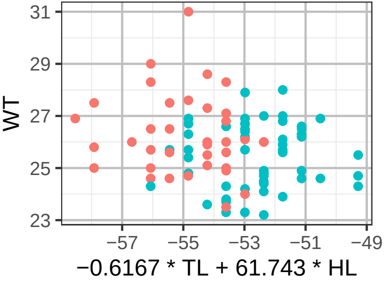
Weight (`WT`) Effect After Accounting for `TL` and `HL`?





Legend:

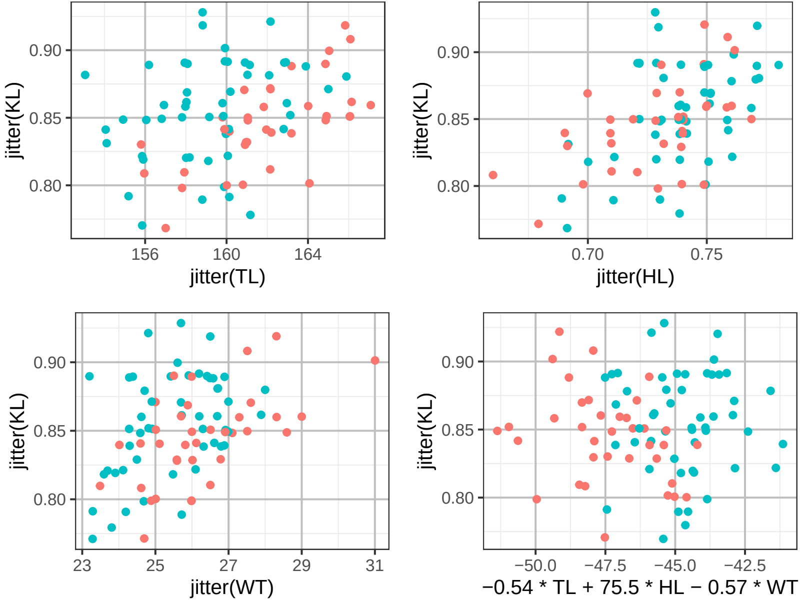
- Red: Perished
- Blue: Survived



Weight (WT) Effect After Accounting for TL and HL?

```
bumpus.fit3 = glm(Survived ~ TL + HL + WT, family=binomial,
                  data=bumpus)
bumpus.fit3$coef
(Intercept)          TL          HL          WT
    46.8813     -0.5435     75.4610     -0.5689
drop1(bumpus.fit3, test="Chisq")
Single term deletions

Model:
Survived ~ TL + HL + WT
       Df Deviance   AIC   LRT   Pr(>Chi)
<none>        75.1  83.1
TL      1     97.3 103.3 22.18 0.00000248
HL      1     99.5 105.5 24.45 0.00000076
WT      1     80.0  86.0  4.93      0.026
```

KL Effect After Accounting for TL, HL, and WT



KL Effect After Accounting for TL, HL, and WT

```
bumpus.fit4 = glm(Survived ~ TL + HL + WT + KL, family=binomial, data=bu
drop1(bumpus.fit4, test="Chisq")
Single term deletions

Model:
Survived ~ TL + HL + WT + KL
       Df Deviance   AIC   LRT   Pr(>Chi)
<none>         68.6  78.6
TL      1      94.7 102.7 26.09 0.00000033
HL      1      86.7  94.7 18.08 0.00002121
WT      1      76.7  84.7  8.10     0.0044
KL      1      75.1  83.1  6.48     0.0109
```

```
bumpus.fit4$coef
(Intercept)          TL          HL          WT          KL
    49.9861     -0.6573     72.3327     -0.7896     27.3775
```

- The coefficients of `TL` and `WT` are negative and of `HL` and `KL` are positive,
- While survivors tended to have lower weight (`WT`) and total length (`TL`) for a given weight and total length, the survivors tended to have larger keels (`KL`) and larger humeruses (`HL`) than the non-survivors.