Lab Worksheet 7

In 1898, Hermon Bumpus, an American biologist working at Brown University, collected data on one of the first examples of natural selection directly observed in nature. Immediately following a bad winter storm, he collected 136 English house sparrows, *Passer domesticus*, and brought them indoors. Of these birds, 64 had died during the storm, but 72 recovered and survived. By comparing measurements of physical traits, Bumpus demonstrated physical differences between the dead and living birds. He interpreted this finding as evidence for natural selection as a result of this storm:

```
bumpus <- read.csv("http://wilkelab.org/classes/SDS348/data_sets/bumpus_full.csv")
head(bumpus)
```

```
##    Sex   Age Survival Length Wingspread Weight Skull_Length Humerus_Length
## 1 Male Adult    Alive    154        241   24.5         31.2           17.4
## 2 Male Adult    Alive    160        252   26.9         30.8           18.7
## 3 Male Adult    Alive    155        243   26.9         30.6           18.6
## 4 Male Adult    Alive    154        245   24.3         31.7           18.8
## 5 Male Adult    Alive    156        247   24.1         31.5           18.2
## 6 Male Adult    Alive    161        253   26.5         31.8           19.8
##   Femur_Length Tarsus_Length Sternum_Length Skull_Width
## 1         17.0          26.0           21.1        14.9
## 2         18.0          30.0           21.4        15.3
## 3         17.9          29.2           21.5        15.3
## 4         17.5          29.1           21.3        14.8
## 5         17.9          28.7           20.9        14.6
## 6         18.9          29.1           22.7        15.4
```

The data set has three categorical variables (`Sex`, with levels `Male` and `Female`, `Age`, with levels `Adult` and `Young`, and `Survival`, with levels `Alive` and `Dead`) and nine numerical variables that hold various aspects of the birds' anatomy, such as wingspread, weight, etc.

Problem 1: *Make a logistic regression model that can predict survival status from all other predictor variables. (Include the categorical predictors `Sex` and `Age`.) Then do backwards selection, removing the predictors with the highest P value one by one, until you are only left with predictors that have P<0.1.*

```
glm.out.complete <- glm(Survival ~ Sex + Age + Length + Wingspread + Weight + Skull_Length + Humerus_
Length + Femur_Length + Tarsus_Length + Sternum_Length + Skull_Width, data=bumpus, family="binomial")
summary(glm.out.complete)
```

```
## 
## Call:
## glm(formula = Survival ~ Sex + Age + Length + Wingspread + Weight + 
##     Skull_Length + Humerus_Length + Femur_Length + Tarsus_Length + 
##     Sternum_Length + Skull_Width, family = "binomial", data = bumpus)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.2342  -0.7890  -0.1887   0.7655   2.1927  
## 
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)    
## (Intercept)    -10.79812   15.13435  -0.713  0.47555    
## SexMale         -1.64710    0.66562  -2.475  0.01334 *  
## AgeYoung         0.32973    0.47216   0.698  0.48496    
## Length           0.42375    0.10958   3.867  0.00011 ***
## Wingspread      -0.01025    0.08496  -0.121  0.90394    
## Weight           0.88472    0.24353   3.633  0.00028 ***
## Skull_Length    -0.46347    0.46141  -1.004  0.31516    
## Humerus_Length  -1.66395    0.89997  -1.849  0.06447 .  
## Femur_Length     0.09391    0.86933   0.108  0.91397    
## Tarsus_Length   -0.25479    0.39646  -0.643  0.52045    
## Sternum_Length  -0.67528    0.32942  -2.050  0.04037 *  
## Skull_Width     -0.68535    0.76052  -0.901  0.36750    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 188.07  on 135  degrees of freedom
## Residual deviance: 129.56  on 124  degrees of freedom
## AIC: 153.56
## 
## Number of Fisher Scoring iterations: 5
```

```
#remove femur length

glm.out <- glm(Survival ~ Sex + Age + Length + Wingspread + Weight + Skull_Length + Humerus_Length +
Tarsus_Length + Sternum_Length + Skull_Width, data=bumpus, family="binomial")
summary(glm.out)
```

```
## 
## Call:
## glm(formula = Survival ~ Sex + Age + Length + Wingspread + Weight + 
##     Skull_Length + Humerus_Length + Tarsus_Length + Sternum_Length + 
##     Skull_Width, family = "binomial", data = bumpus)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.2444  -0.7987  -0.1872   0.7588   2.1838  
## 
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)    
## (Intercept)    -10.90876   15.08451  -0.723 0.469571    
## SexMale         -1.65241    0.66405  -2.488 0.012833 *  
## AgeYoung         0.32606    0.47066   0.693 0.488453    
## Length           0.42440    0.10950   3.876 0.000106 ***
## Wingspread      -0.01035    0.08493  -0.122 0.903042    
## Weight           0.88092    0.24042   3.664 0.000248 ***
## Skull_Length    -0.45503    0.45461  -1.001 0.316864    
## Humerus_Length  -1.61244    0.76223  -2.115 0.034393 *  
## Tarsus_Length   -0.23454    0.34926  -0.672 0.501877    
## Sternum_Length  -0.67692    0.32915  -2.057 0.039727 *  
## Skull_Width     -0.68140    0.75958  -0.897 0.369679    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 188.07  on 135  degrees of freedom
## Residual deviance: 129.57  on 125  degrees of freedom
## AIC: 151.57
## 
## Number of Fisher Scoring iterations: 5
```

```
#Wingspread
glm.out <- glm(Survival ~ Sex + Age + Length +  Weight + Skull_Length + Humerus_Length + Tarsus_Length + Sternum_Length + Skull_Width, data=bumpus, family="binomial")
summary(glm.out)
```

```
## 
## Call:
## glm(formula = Survival ~ Sex + Age + Length + Weight + Skull_Length + 
##     Humerus_Length + Tarsus_Length + Sternum_Length + Skull_Width, 
##     family = "binomial", data = bumpus)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.2472  -0.7907  -0.1847   0.7602   2.1922  
## 
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)    
## (Intercept)    -11.7689    13.3246  -0.883 0.377102    
## SexMale         -1.6949     0.5657  -2.996 0.002736 ** 
## AgeYoung         0.3355     0.4641   0.723 0.469700    
## Length           0.4197     0.1023   4.102 4.09e-05 ***
## Weight           0.8799     0.2402   3.663 0.000249 ***
## Skull_Length    -0.4491     0.4517  -0.994 0.320046    
## Humerus_Length  -1.6458     0.7111  -2.315 0.020637 *  
## Tarsus_Length   -0.2429     0.3424  -0.709 0.478129    
## Sternum_Length  -0.6833     0.3251  -2.102 0.035534 *  
## Skull_Width     -0.6853     0.7583  -0.904 0.366162    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 188.07  on 135  degrees of freedom
## Residual deviance: 129.59  on 126  degrees of freedom
## AIC: 149.59
## 
## Number of Fisher Scoring iterations: 5
```

```
#tarsus length
glm.out <- glm(Survival ~ Sex + Age + Length +  Weight + Skull_Length + Humerus_Length  + Sternum_Len
gth + Skull_Width, data=bumpus, family="binomial")
summary(glm.out)
```

```
## 
## Call:
## glm(formula = Survival ~ Sex + Age + Length + Weight + Skull_Length + 
##     Humerus_Length + Sternum_Length + Skull_Width, family = "binomial", 
##     data = bumpus)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.2465  -0.8113  -0.1847   0.7575   2.1017  
## 
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)    
## (Intercept)    -10.7332    13.1537  -0.816 0.414509    
## SexMale         -1.5723     0.5368  -2.929 0.003403 ** 
## AgeYoung         0.3565     0.4624   0.771 0.440738    
## Length           0.4173     0.1012   4.122 3.76e-05 ***
## Weight           0.8720     0.2393   3.645 0.000268 ***
## Skull_Length    -0.5098     0.4416  -1.154 0.248299    
## Humerus_Length  -1.9455     0.5887  -3.305 0.000950 ***
## Sternum_Length  -0.6725     0.3235  -2.079 0.037597 *  
## Skull_Width     -0.7033     0.7561  -0.930 0.352291    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 188.07  on 135  degrees of freedom
## Residual deviance: 130.08  on 127  degrees of freedom
## AIC: 148.08
## 
## Number of Fisher Scoring iterations: 5
```

```
#Age
glm.out <- glm(Survival ~ Sex + Length +  Weight + Skull_Length + Humerus_Length +  Sternum_Length + 
Skull_Width, data=bumpus, family="binomial")
summary(glm.out)
```

```
## 
## Call:
## glm(formula = Survival ~ Sex + Length + Weight + Skull_Length + 
##     Humerus_Length + Sternum_Length + Skull_Width, family = "binomial", 
##     data = bumpus)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.2938  -0.7823  -0.1952   0.7758   2.0455  
## 
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)    
## (Intercept)    -11.2855    12.9825  -0.869 0.384691    
## SexMale         -1.5916     0.5324  -2.990 0.002794 ** 
## Length           0.4220     0.1009   4.182 2.89e-05 ***
## Weight           0.8556     0.2350   3.641 0.000272 ***
## Skull_Length    -0.5374     0.4378  -1.228 0.219594    
## Humerus_Length  -1.9022     0.5836  -3.259 0.001117 ** 
## Sternum_Length  -0.6851     0.3207  -2.136 0.032643 *  
## Skull_Width     -0.6582     0.7511  -0.876 0.380807    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 188.07  on 135  degrees of freedom
## Residual deviance: 130.68  on 128  degrees of freedom
## AIC: 146.68
## 
## Number of Fisher Scoring iterations: 5
```

```
#Skull Width
glm.out <- glm(Survival ~ Sex + Length +  Weight + Skull_Length + Humerus_Length +  Sternum_Length ,
data=bumpus, family="binomial")
summary(glm.out)
```

```
##
## Call:
## glm(formula = Survival ~ Sex + Length + Weight + Skull_Length +
##     Humerus_Length + Sternum_Length, family = "binomial", data = bumpus)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4048  -0.7911  -0.1888   0.7747   1.9636
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -15.6788    11.9179  -1.316 0.188319
## SexMale         -1.5427     0.5285  -2.919 0.003511 **
## Length           0.4193     0.1011   4.149 3.34e-05 ***
## Weight           0.8319     0.2321   3.584 0.000339 ***
## Skull_Length    -0.6294     0.4228  -1.488 0.136635
## Humerus_Length  -1.9684     0.5770  -3.412 0.000646 ***
## Sternum_Length  -0.7108     0.3205  -2.218 0.026585 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 188.07  on 135  degrees of freedom
## Residual deviance: 131.46  on 129  degrees of freedom
## AIC: 145.46
##
## Number of Fisher Scoring iterations: 5
```

```
#Skull Length
glm.out <- glm(Survival ~ Sex + Length +  Weight  + Humerus_Length +  Sternum_Length , data=bumpus, f
amily="binomial")
summary(glm.out)
```

```
##
## Call:
## glm(formula = Survival ~ Sex + Length + Weight + Humerus_Length +
##     Sternum_Length, family = "binomial", data = bumpus)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4921  -0.7678  -0.2155   0.7890   2.0192
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -23.15186   10.83789  -2.136 0.032663 *
## SexMale         -1.39306    0.51054  -2.729 0.006360 **
## Length           0.38266    0.09487   4.034 5.49e-05 ***
## Weight           0.76098    0.22248   3.420 0.000625 ***
## Humerus_Length  -2.17650    0.55596  -3.915 9.05e-05 ***
## Sternum_Length  -0.75484    0.31296  -2.412 0.015870 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 188.07  on 135  degrees of freedom
## Residual deviance: 133.72  on 130  degrees of freedom
## AIC: 145.72
##
## Number of Fisher Scoring iterations: 5
```

Problem 2: *Using a cut-off of 0.6, what is the true positive rate and true negative rate for the final model created in Problem 1?*

```
cutoff <- 0.6
# Make a data frame with proabilities and true survival outcomes
pred_data <- data.frame(probability=glm.out$fitted.values, Survival=bumpus$Survival)

# Count the number of alive birds correctly identified as alive (true positive)
pred_data %>% filter(probability < cutoff & Survival=="Alive") %>%
  tally() -> true_pos

# Count the number of dead birds correctly identified as dead (true negative)
pred_data %>% filter(probability >= cutoff & Survival=="Dead") %>%
  tally() -> true_neg

# Count the total number of alive birds (known positives)
pred_data %>% filter(Survival=="Alive") %>%
  tally() -> pos_total

# Count the total number of dead birds (known negatives)
pred_data %>% filter(Survival=="Dead") %>%
  tally() -> neg_total

true_pos_rate <- true_pos$n/pos_total$n
true_neg_rate <- true_neg$n/neg_total$n

true_pos_rate
```

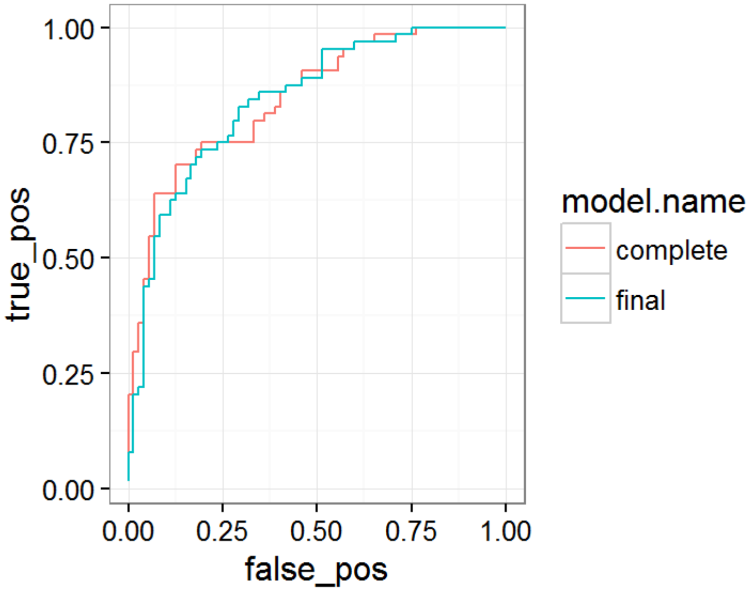
```
## [1] 0.875
```

Problem 3: *Make ROC curves for the complete model (using all predictors) and the final, selected model (using only predictors with P<0.1) from Question 1 and plot them jointly in one figure. Use the function `calc_ROC()` given below. How do the two ROC curves differ?*

```r
calc_ROC <- function(probabilities, known_truth, model.name=NULL)
  {
  outcome <- as.numeric(factor(known_truth))-1
  pos <- sum(outcome) # total known positives
  neg <- sum(1-outcome) # total known negatives
  pos_probs <- outcome*probabilities # probabilities for known positives
  neg_probs <- (1-outcome)*probabilities # probabilities for known negatives
  true_pos <- sapply(probabilities,
                     function(x) sum(pos_probs>=x)/pos) # true pos. rate
  false_pos <- sapply(probabilities,
                     function(x) sum(neg_probs>=x)/neg)
  if (is.null(model.name))
    result <- data.frame(true_pos, false_pos)
  else
    result <- data.frame(true_pos, false_pos, model.name)
  result %>% arrange(false_pos, true_pos)
  }

ROC.complete <- calc_ROC(probabilities=glm.out.complete$fitted.values,
                 known_truth=bumpus$Survival,
                 model.name="complete")
ROC.final <- calc_ROC(probabilities=glm.out$fitted.values,
                 known_truth=bumpus$Survival,
                 model.name="final")

ROCs <- rbind(ROC.complete, ROC.final)
ggplot(data=ROCs, aes(x=false_pos, y=true_pos, color=model.name)) +
  geom_line()
```



There is barely any difference between the two ROC curves. The complete model and the final, selected model perform approximately the same.

Problem 4: *Calculate AUCs for the complete model (using all parameters) and the final, selected model from Problem 1. What do you find? Do the AUCs support your conclusions from Problem 2?*

```
ROCs%>% group_by(model.name) %>%
  mutate(delta=false_pos-lag(false_pos)) %>%
  summarize(AUC=sum(delta*true_pos, na.rm=T)) %>%
  arrange(desc(AUC))
```

```
## Source: local data frame [2 x 2]
##
##   model.name       AUC
##       (fctr)     (dbl)
## 1   complete 0.8463542
## 2      final 0.8422309
```

Yes the area under the curves are similar in value meaning they are basically the same.