Homework 6

Manav Mandhani, mm58926

This homework is due on Mar 10, 2015 in class.

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.

Question 1 (5 pts): 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. How many and which predictors remain in the final model?

```
glm.out <- glm(Survival ~ Sex + Length + Weight +  Humerus_Length + Sternum_Length,
               data = bumpus,
               family = binomial) # family = binomial required for logistic regression
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
```

We are left with 5 predictors: Sex, Length, Weight, Humerus_Length and Sternum_Length

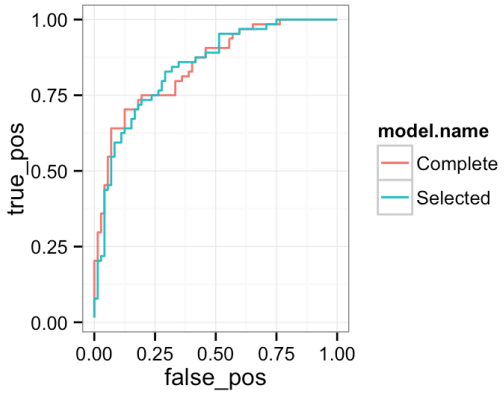
Question 2 (2 pt): Make ROC curves for the complete model (using all predictors) and the final, selected model (using only predictors with P<0.1) from Question 1 and plot them jointly in one figure. Use the function `calc_ROC()` *given below. How do the two ROC curves differ?*

```
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)
  }

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

```
## Warning in model.matrix.default(mt, mf, contrasts): the response appeared
## on the right-hand side and was dropped
```

```
## Warning in model.matrix.default(mt, mf, contrasts): problem with term 3 in
## model.matrix: no columns are assigned
```

```
ROC1 <- calc_ROC(probabilities=glm.out$fitted.values, # predicted probabilities
                 known_truth=bumpus$Survival,      # the known truth, i.e., true species assignment
                 model.name="Complete")

glm.out <- glm(Survival ~ Sex + Length + Weight +  Humerus_Length + Sternum_Length,
               data = bumpus,
               family = binomial) # family = binomial required for logistic regression

ROC2 <- calc_ROC(probabilities=glm.out$fitted.values, # predicted probabilities
                 known_truth=bumpus$Survival,      # the known truth, i.e., true species assignment
                 model.name="Selected")

ggplot(data=NULL, aes(x=false_pos, y=true_pos)) +
  geom_line(data=ROC1, aes(color=model.name)) +
  geom_line(data=ROC2, aes(color=model.name))
```



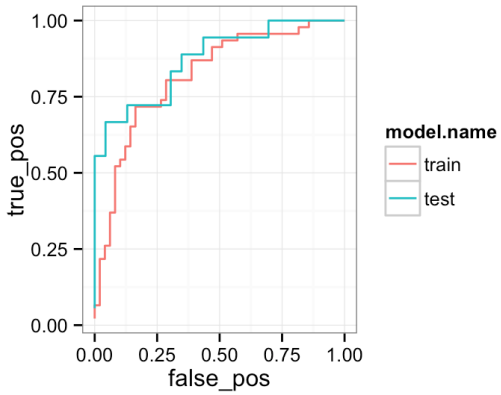
The selected ROC curve has a higher true positive rate, giving it a slightly higher area under the ROC curve.

Question 3 (3 pt): Now split the data set into 70% training data and 30% test data, fit the final model on the training data set, and then evaluate the performance of the model on the test data set by calculating the area under the ROC curve for the test and the training data set. Adapt the code from the class 13 worksheet do accomplish this. What do you find?

```
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)
}

train_fraction <- 0.7 # fraction of data for training purposes
set.seed(101)  # set the seed to make your partition reproductible
n_obs <- nrow(bumpus) # number of observations in biopsy data set
train_size <- floor(train_fraction * nrow(bumpus)) # number of observations in training set
train_indices <- sample(1:n_obs, size = train_size)

train_data <- bumpus[train_indices, ] # get training data
test_data <- bumpus[-train_indices, ] # get test data

glm.out.train <- glm(Survival ~ Sex + Length + Weight +  Humerus_Length + Sternum_Length,
                data=train_data,
                family=binomial)
head(train_data)
```

```
##       Sex   Age Survival Length Wingspread Weight Skull_Length
## 51   Male Adult     Dead    161        244   25.0         31.3
## 6    Male Adult    Alive    161        253   26.5         31.8
## 96 Female Adult    Alive    164        248   24.2         32.7
## 88 Female Adult    Alive    156        245   25.3         31.6
## 33   Male Adult    Alive    160        247   24.6         32.3
## 40   Male Adult     Dead    163        250   25.5         32.5
##    Humerus_Length Femur_Length Tarsus_Length Sternum_Length Skull_Width
## 51           17.8         17.4          27.5           22.2        15.1
## 6            19.8         18.9          29.1           22.7        15.4
## 96           19.1         19.1          30.5           21.1        15.3
## 88           18.5         18.0          29.3           20.5        15.7
## 33           19.2         18.9          28.8           23.0        15.4
## 40           19.1         18.6          30.4           22.6        15.8
```

```
test_pred <- predict(glm.out.train, test_data, type='response')
ROC.train <- calc_ROC(probabilities=glm.out.train$fitted.values,
                known_truth=train_data$Survival,
                model.name="train")
ROC.test <- calc_ROC(probabilities=test_pred,
                known_truth=test_data$Survival,
                model.name="test")
ROCs <- rbind(ROC.train, ROC.test)
ggplot(ROCs, aes(x=false_pos, y=true_pos, color=model.name)) +
  geom_line()
```




```
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
## 1       test 0.8719807
## 2      train 0.8185448
```

The test data performs better than the training data, as evident by it's higher AUC.