CMDA-3654

Homework 7

Michael La Vance

Due as a .pdf upload

Problem 1: [50 pts] Logistic Regression

Hermon Bumpus analyzed various characteristics of some house sparrows that were found on the ground after a severe winter storm in 1898. Some of the sparrows survived and some perished. The data on male sparrows in found in `bumpus.csv` are survival status (`survived`, `perished`), age (`1 = adult`, `2 = juvenile`), the total length from tip of beak to tip of tail (in mm), the alar extent (length from tip to tip of the extended wings, in mm), the weight in grams, the length of the head in mm, the length of the humerus (arm bone, in inches), the length of the femur (thigh bones, in inches), the length of the tibio-tarsus (leg bone, in inches), the breadth of the skull in inches, and the length of the sternum in inches.

Analyze the data to see whether the probability of survival is associated with physical characteristics of the birds.

This would be consistent, according to Bumpus, with the theory of natural selection: those that survived did so because of some superior physical traits. Realize that (i) the sampling is from a population of grounded sparrows, and (ii) physical measurements and survival are both random.

a. Assuming that Weight is the only explanatory variable, fit a logistic regression model with Status as the response variable and answer the following questions.

   i. What is the probability a bird that weighs 25 grams survives?

   ii. What is the probability a bird that weighs 30 grams survives?

   iii. Plot the logistic regression model with a scatterplot of the observations. Make sure all plot elements make sense.

   iv. Suppose we come up with a classification rule that says we will consider a bird as having survived if the probability of surviving is 60% or greater. For what body weights would this be associated with?

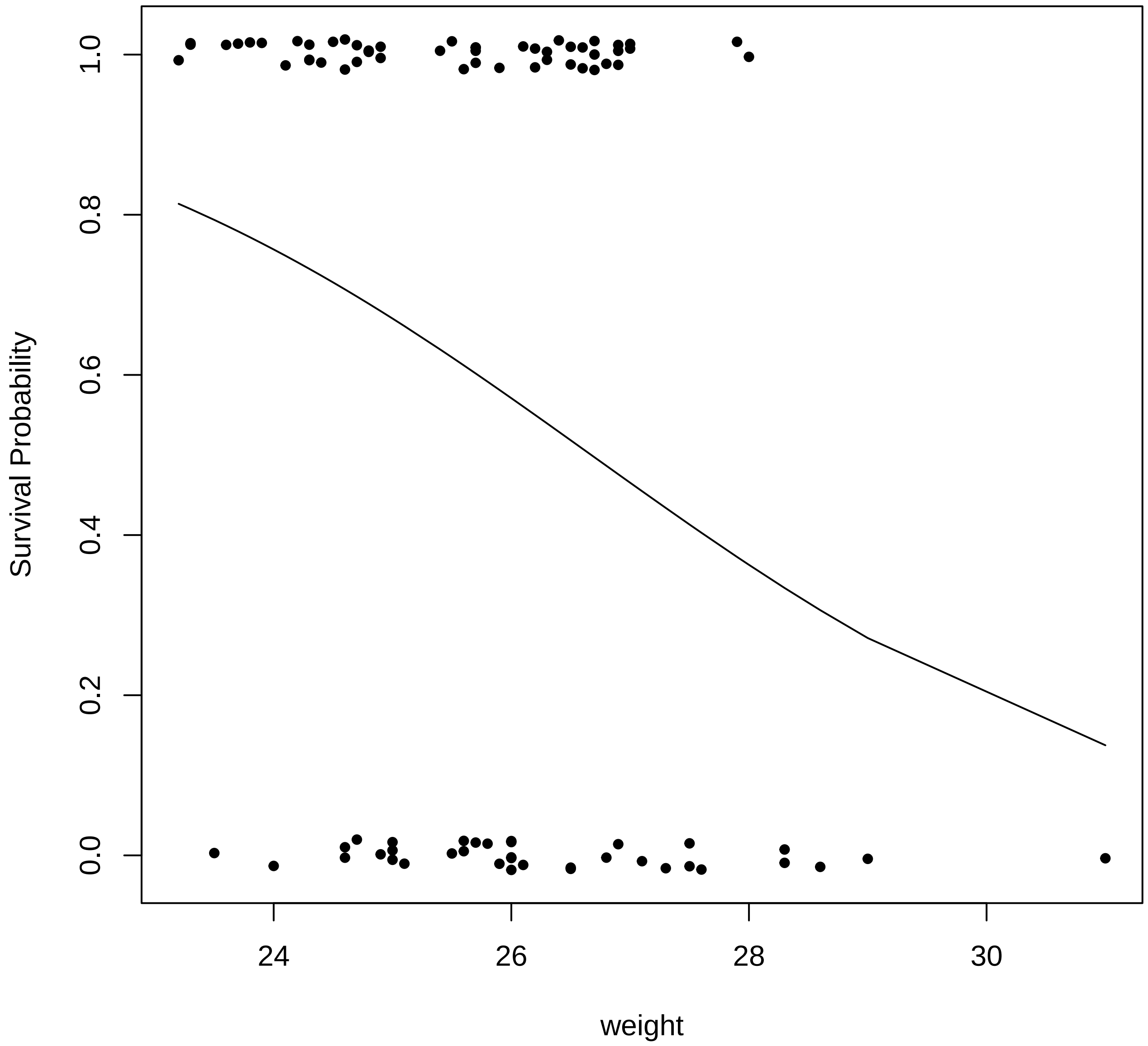
```
bumpus <- read.csv("bumpus.csv")
#changing age to Boolean
# 0 = child, 1 = adult
bumpus$age <- bumpus$age - 1
#changing survival status to boolean
# 0 = dead, 1 = survived
bumpus <- bumpus %>%
  mutate(status = if_else(status == "Survived", 1, 0))
bumpusglm <- glm(status ~ weight, data = bumpus, family = binomial)



pi.hat.ordered <- predict(bumpusglm, data.frame(weight = sort(bumpus$weight)), type = "response")

#plotting
plot(jitter(status, amount =.02)~weight, data = bumpus, pch = 20, ylab = "Survival Probability")


lines(pi.hat.ordered ~ sort(weight), data= bumpus)
```



If we said all birds with a probability of survival of 60% or higher surved, all birds under 25.72g would suvive.

b. Now consider using all of the physical characteristics as possible predictor variables in a logistic regression with Status as the response. Find the best subset of explanatory variables using `stepAIC()`. State the best model in terms of log-odds. Use this model for the remaining questions.

```
# This full model has three predictor variables
modfit_full <- glm(status ~ . , data = bumpus, family = binomial)
# The null model has no predictor variables
modfit_null <- glm(status ~ 1, data = bumpus, family = binomial)

modfit_best <- stepAIC(modfit_full, scope = list(lower = modfit_null, modfit_full),
                       direction = "both", trace = 0)
coefficients(modfit_best)
```

```
   (Intercept)   total_length         weight humerus_length        sternum
    49.9860907     -0.6573497     -0.7895756     72.3326762     27.3774705
```

The log odds of survival suggested by this data set are 49.99 -.657x total_length -.789x weight + 72.3x humerus length +

27.4x sternum length.

c. Is age group important? If so, how does the odds of survival change?

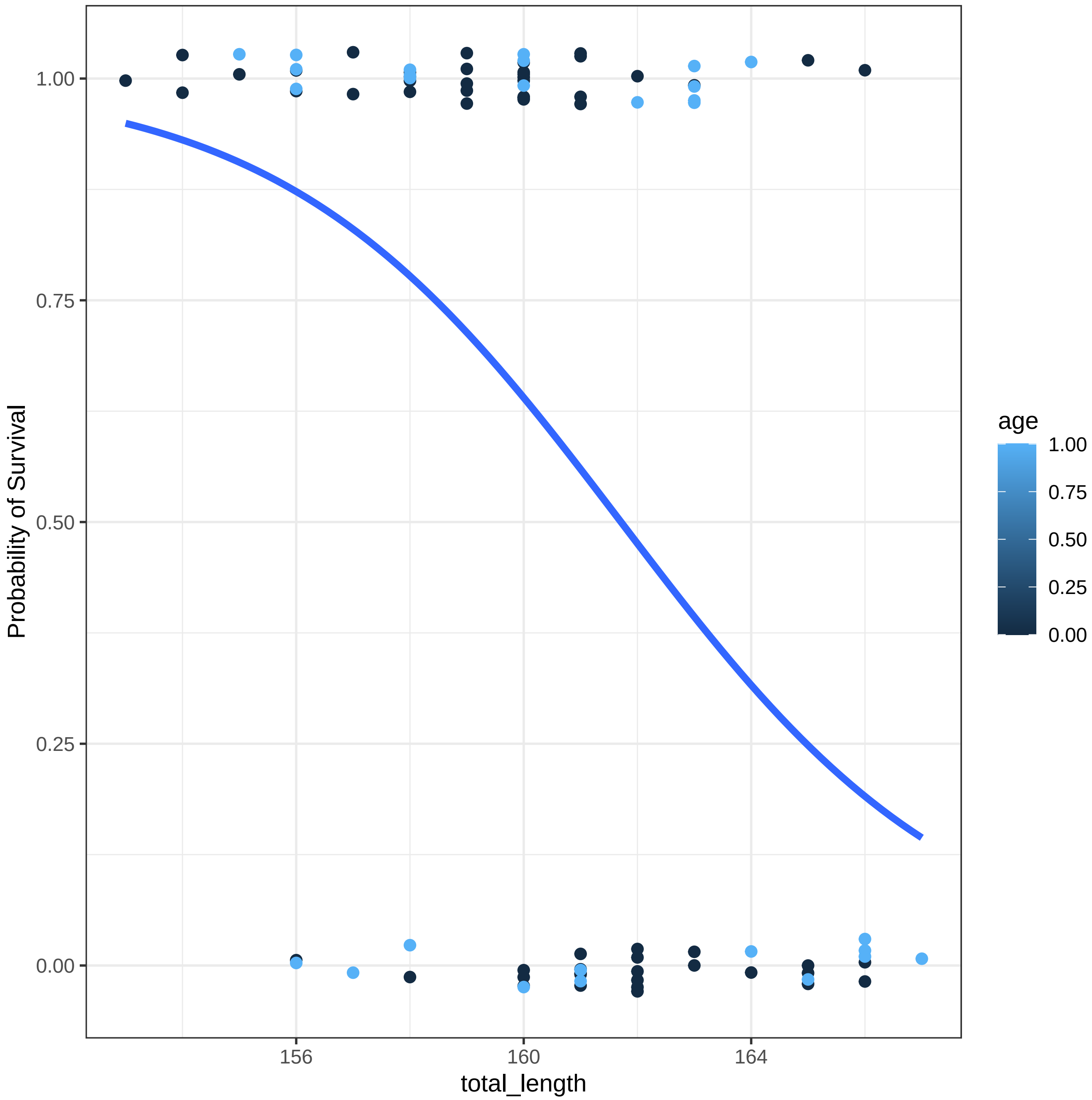
The age might be important, but as our data regarding age is only if they are a child or not, we can not determine if it is important with our current data.

d. Is total length important? If so, if the total length is increased from 160 to 165 mm, and assuming everything else is held constant, what is change in odds of survival?

If everything except total length was held constant, our model says that an increase in 5mm to total length would decrease the log odds of survival by 3.286.

e. Plot Status versus η the log-odds function and overlay the logistic regression curve.

```
ggplot(bumpus, aes(x = total_length, y = as.numeric(status))) +
  geom_point(aes(color = age), position = position_jitter(height = 0.03, width = 0), size = 2) +
  geom_smooth(aes(color = age), method = "glm", method.args = list(family = "binomial"), se = F, size = 1.5) +
  labs(y = "Probability of Survival") + theme_bw()
```



\end{enumerate}

Problem 2: [50 pts] Classification using LDA, QDA, and SVM

Load the `wine` dataset from the `rattle` package in R.

Consider `Type` to be the response variable, and all other variables as features.

a. Describe the dataset in your own words, in 2-3 lines.

The wine dataset is a data set with 178 samples with 13 variables, all numeric and continuous.The wine type is classified into one of 3 types each taking a value, one, two, or three.

b. Perform classification using LDA (linear discriminant analysis). Display the Confusion Matrix. Report the classification error rate.

```
lda.wine <- lda(Type ~ . , data = wine)

pred.wine <- predict(lda.wine, data = wine)


#calculating error rate

sum(wine$Type != pred.wine$class)/nrow(wine)
```

```
[1] 0
```

```
#confusion matrix
confusionMatrix(pred.wine$class, wine$Type)$table
```

```
          Reference
Prediction  1  2  3
         1 59  0  0
         2  0 71  0
         3  0  0 48
```

This has 100% classification rate.

c. Perform classification using QDA (quadratic discriminant analysis). Display the Confusion Matrix. Report the classification error rate.

```
qda.wine <- qda(Type ~ ., data = wine)


qda.wine.conf.mat <- confusionMatrix(predict(qda.wine)$class, wine$Type)
qda.wine.conf.mat$table
```

```
          Reference
Prediction  1  2  3
         1 59  1  0
         2  0 70  0
         3  0  0 48
```

99.44% accurate.

d. Perform classification using SVM (support vector machines). Display the Confusion Matrix. Report the classification error rate.

```
library(e1071)

svm.wine <- svm(Type ~ ., data = wine)

pred.wine<- predict(svm.wine, data = wine)
```

```
confusionMatrix(pred.wine, wine$Type)$table
```

```
          Reference
Prediction  1  2  3
         1 59  0  0
         2  0 71  0
         3  0  0 48
```

```
mean(wine$Type != pred.wine)
```

```
[1] 0
```

100% accurate

e. Rank the classification methods in your order of preference for this dataset, and justify your preference. (Hint: Note that the error rates should be calculated by cross-validation)

```
#from lecture
train.control <- trainControl(method = "cv", number = 10)

model.lda <- train(Type ~ ., data = wine, method = "lda", trControl = train.control)
model.qda <- train(Type ~ ., data = wine, method = "qda", trControl = train.control)
model.svm <- train(Type ~ ., data = wine, method = "svmLinear", trControl = train.control)

train.control.loocv <- trainControl(method = "LOOCV")


model.lda.loocv <- train(Type ~ ., data = wine, method = "lda", trControl = train.control.loocv)
model.qda.loocv <- train(Type ~ ., data = wine, method = "qda", trControl = train.control.loocv)
model.svm.loocv <- train(Type ~ ., data = wine, method = "svmLinear", trControl = train.control.loocv)


model.lda$results$Accuracy
```

```
[1] 0.9830065
```

```
model.qda$results$Accuracy
```

```
[1] 0.9944444
```

```
model.svm$results$Accuracy
```

```
[1] 0.9555556
```

rank them from best to worst, quadratic, linear, support vector. I think that the support vector can be too tailored to the training data and lose overall accuracy, while the linear one can not be over specified I also think that has a hard time generating accurate boundaries as the variation between the data increases. The quadratic method is still hurt by large variation like the linear case, but not to as much of an extent.

Problem 3: [20 pts Extra Credit]: More Logistic Regression

Load the `mtcars` data in R.

a. Describe the variable `am` in one sentence. **We will consider `am` to be the response variable in the following questions.**

b. Construct a plot of hp (x-axis) and wt (y-axis), with different colors for automatic and manual transmission. From the plot, do you think automatic and manual transmission can be distinguished by weight and horsepower?

c. Fit a logistic regression model with `wt` as the only feature. Using this model, explain whether heavier cars are more likely or less likely to have manual transmission. If weight increases by 1000 lbs, what is the change in odds of a car having manual transmission?

d. Fit a logistic regression model with `hp` as the only feature. Using this model, explain whether cars with higher horsepower are more likely or less likely to have manual transmission. If horsepower increases by 100, what is the change in odds of a car having manual transmission?

e. If you had to choose between these two models, which one would you choose and why?