CMDA-3654

Homework 7

Eduardo Salvador

Due as a .pdf upload

Problem 1: [50 pts] Logistic Regression

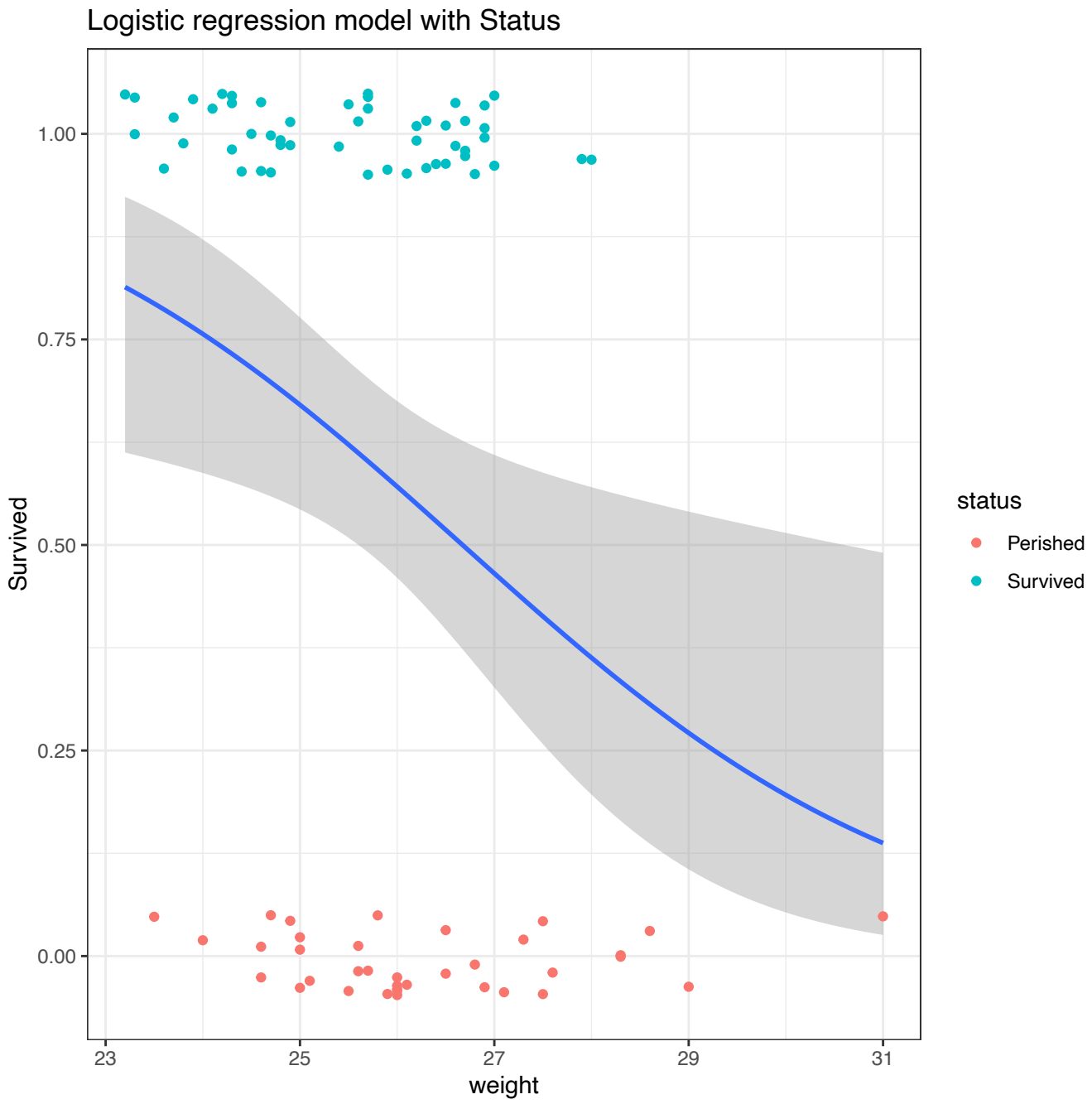
Hermon Bumpus analyzed various characteristics of some house sparrows that were found on the ground after a severe winter storm in 1898. Some of the sparrows survived and some perished. The data on male sparrows in found in `bumpus.csv` are survival status (`survived`, `perished`), age (`1 = adult`, `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.

```
library(tidyverse)
#Reading csv file
bump<-read.csv("/Users/eduardosalvador/Desktop/FINAL\ Spring\ Semester\ 2021/CMDA\ /Assignments/HW7/bumpus.csv
#Creating y variable of people who survived and those who have perished
bump <-bump %>%
  mutate(mystatus=if_else(status=="Survived",1,0))
#Creating regression model with ggplot
ggplot(bump,aes(x=weight,y=mystatus))+theme_bw()+ylab("Survived")+ggtitle("Logistic regression model with Stat
  geom_smooth(method="glm",method.args=list(family="binomial"))+geom_point(aes(color=status),position=position
```



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

```
#Creating fitting generalized linear model
bumpglm<-glm(mystatus~weight,data=bump,family=binomial)
head(bumpglm)
```

```
$coefficients
(Intercept)      weight
 11.3201190  -0.4244016

$residuals
        1         2         3         4         5         6         7         8
 1.397612  2.101072  2.101072  1.365255  1.335531  1.929158  1.414850  1.350078
        9        10        11        12        13        14        15        16
 1.271379  1.818079  1.818079  1.451600  1.582563  1.283144  1.661664  1.661664
       17        18        19        20        21        22        23        24
```

```
 1.929158  2.011470  1.308227  1.432835  2.756154  2.683182  1.720279  1.661664
       25        26        27        28        29        30        31        32
 1.969441  1.229007  1.661664  1.853546  1.365255  2.011470  1.471178  1.295419
       33        34        35        36        37        38        39        40
 1.634170  2.148808  1.432835 -2.076243 -2.275370 -2.576863 -2.388351 -2.645226
       41        42        43        44        45        46        47        48
-1.674782 -2.448541 -3.122339 -2.330662 -2.076243 -2.330662 -1.834298 -2.949627
       49        50        51        52        53        54        55        56
-2.330662 -2.576863 -3.034151 -3.410512 -3.034151 -2.330662 -1.501352 -3.410512
       57        58        59        60        61        62        63        64
-1.704036 -1.159401 -1.501352  1.414850  1.607819  1.451600  1.853546  1.381090
       65        66        67        68        69        70        71        72
 1.238936  2.011470  1.890550  2.101072  1.365255  2.148808  2.055320  1.471178
       73        74        75        76        77        78        79        80
 1.784086  1.969441  1.238936 -4.109556 -1.947579 -4.844644 -1.908205 -1.441416
       81        82        83        84        85        86        87
-3.310350 -1.766404 -2.511341 -1.372496 -3.034151 -1.704036 -2.330662

$fitted.values
        1         2         3         4         5         6         7         8
0.7155063 0.4759474 0.4759474 0.7324639 0.7487657 0.5183607 0.7067889 0.7406980
        9        10        11        12        13        14        15        16
0.7865478 0.5500311 0.5500311 0.6888951 0.6318863 0.7793359 0.6018064 0.6018064
       17        18        19        20        21        22        23        24
0.5183607 0.4971490 0.7643936 0.6979171 0.3628244 0.3726918 0.5813011 0.6018064
       25        26        27        28        29        30        31        32
0.5077583 0.8136648 0.6018064 0.5395066 0.7324639 0.4971490 0.6797272 0.7719510
       33        34        35        36        37        38        39        40
0.6119313 0.4653743 0.6979171 0.5183607 0.5605111 0.6119313 0.5813011 0.6219604
       41        42        43        44        45        46        47        48
0.4029073 0.5915936 0.6797272 0.5709374 0.5183607 0.5709374 0.4548322 0.6609741
       49        50        51        52        53        54        55        56
0.5709374 0.6119313 0.6704185 0.7067889 0.6704185 0.5709374 0.3339337 0.7067889
       57        58        59        60        61        62        63        64
0.4131580 0.1374858 0.3339337 0.7067889 0.6219604 0.6888951 0.5395066 0.7240657
       65        66        67        68        69        70        71        72
0.8071445 0.4971490 0.5289466 0.4759474 0.7324639 0.4653743 0.4865421 0.6797272
       73        74        75        76        77        78        79        80
0.5605111 0.5077583 0.8071445 0.7566647 0.4865421 0.7935865 0.4759474 0.3062378
       81        82        83        84        85        86        87
0.6979171 0.4338782 0.6018064 0.2714006 0.6704185 0.4131580 0.5709374

$effects
(Intercept)      weight
 -1.4772073   2.4414522   0.8957280   0.6474331   0.6349562   0.8413825

  0.6671153   0.6411371   0.6054655   0.8057024   0.8057024   0.6810123

  0.7275775   0.6112022   0.7543449   0.7543449   0.8413825   0.8675212

  0.6229023   0.6739782   1.0949774   1.0733857   0.7737856   0.7543449

  0.8542034   0.5831223   0.7543449   0.8171517   0.6474331   0.8675212

  0.6882306   0.6170120   0.7451209   0.9106571   0.6739782  -1.1599673

 -1.2201382  -1.3069592  -1.2531900  -1.3260869  -1.0257537  -1.2705325

 -1.4550180  -1.2363991  -1.1599673  -1.2363991  -1.0820472  -1.4091991
```

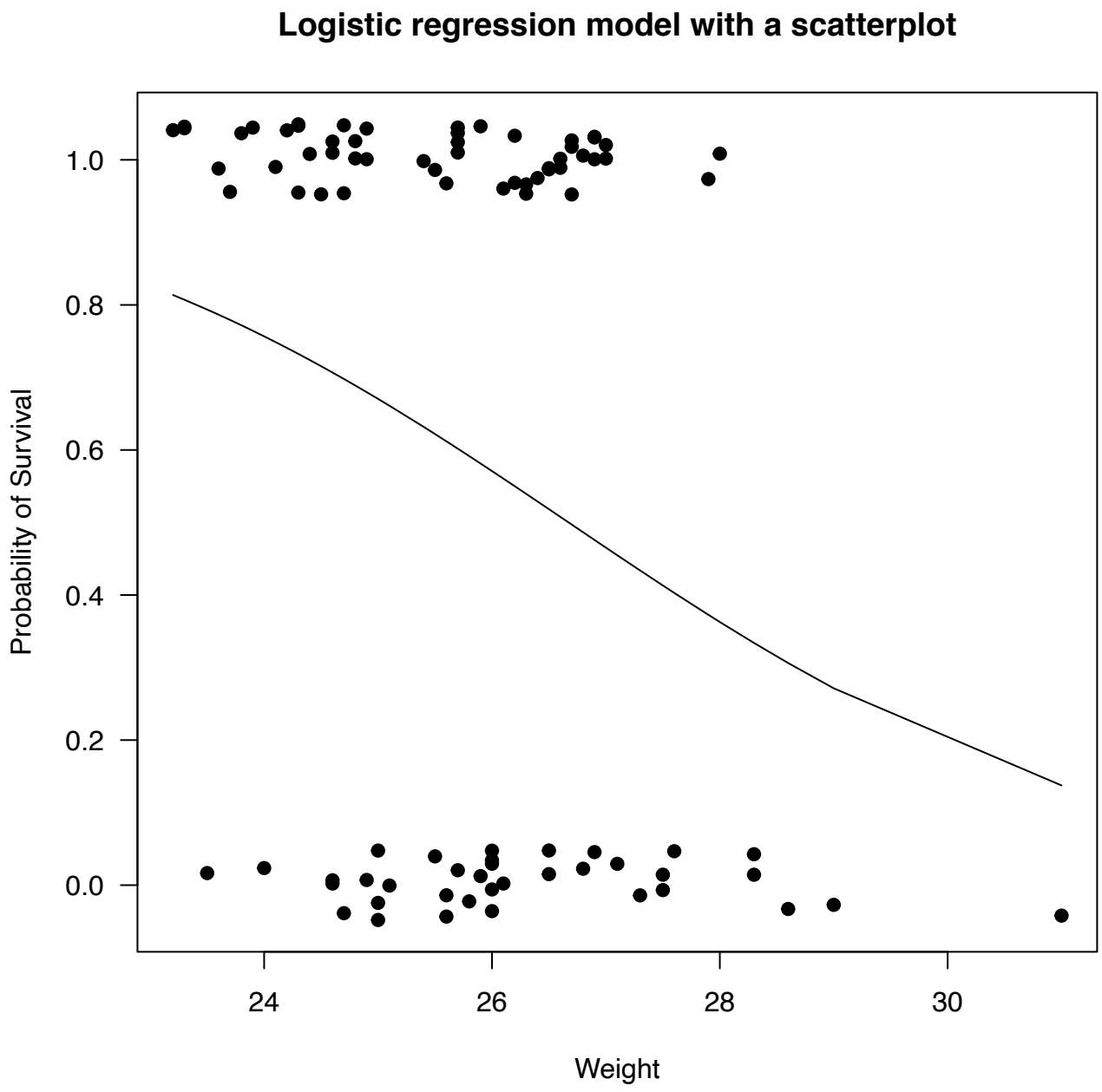
```
 -1.2363991  -1.3069592  -1.4317355  -1.5295544  -1.4317355  -1.2363991

 -0.9559097  -1.5295544  -1.0365033  -0.7249919  -0.9559097   0.6671153

  0.7362041   0.6810123   0.8171517   0.6538543   0.5886292   0.8675212

  0.8290386   0.8957280   0.6474331   0.9106571   0.8813561   0.6882306

  0.7946710   0.8542034   0.5886292  -1.7015995  -1.1193688  -1.8709810

 -1.1065901  -0.9281538  -1.5039064  -1.0587324  -1.2884483  -0.8924854

 -1.4317355  -1.0365033  -1.2363991

$R
            (Intercept)      weight
(Intercept)   -4.417004 -114.334632
weight         0.000000   -5.752693

$rank
[1] 2


summary(bumpglm)

Call:
glm(formula = mystatus ~ weight, family = binomial, data = bump)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7764  -1.2455   0.7607   1.0247   1.4240

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  11.3201     4.5053   2.513   0.0120 *
weight       -0.4244     0.1738  -2.441   0.0146 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 118.01  on 86  degrees of freedom
Residual deviance: 111.25  on 85  degrees of freedom
AIC: 115.25

Number of Fisher Scoring iterations: 4

#Use predict function to get probability a bird that weights 25 grams
predict(bumpglm,newdata = data.frame(weight=25),type="response")

        1
0.6704185

#The probability a bird that weighs 25 grams survives is 67%.

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

#Use predict function to get probability a bird that weights 30 grams
predict(bumpglm,newdata = data.frame(weight=30),type="response")
```

```
       1
0.1959299
```

```
#The probability a bird that weighs 30 grams survives is 19.59%.
```

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

```
#Creating scatter points with plot
plot(jitter(mystatus, amount = 0.05)~weight,data=bump,pch=19,las=1, main="Logistic regression model with a sca
#Creating line for regression model with lines function
ordw<-predict(bumpglm,data.frame(weight=sort(bump$weight)),type = "response")
lines(ordw~sort(weight),data=bump)
```



iv. Suppose we come up with a classification rule that says we will consider a bird as having survived if the

```
#The body weights would have to be 25.6 or lower.
```

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.

```
library (MASS)
#Creating fitting generalized linear model with all the variables
full_model<-glm(mystatus~age+total_length+alar_extent+weight+head_length+humerus_length+femur_length+tibio_tar
#Creating fitting generalized linear model with null
null_model <- glm(mystatus ~ 1, data = bump, family= binomial)
#Finding best best subset of exploratory values using stepAIC()
best_model <- stepAIC(full_model, scope = list(lower = null_model, full_model),trace = 0, direction = "both")
summary (best_model)
```

```
Call:
glm(formula = mystatus ~ total_length + weight + humerus_length +
    sternum, family = binomial, data = bump)

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 **
total_length    -0.6573     0.1683  -3.907 9.35e-05 ***
weight          -0.7896     0.3097  -2.549 0.010800 *
humerus_length  72.3327    20.7640   3.484 0.000495 ***
sternum         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
```

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

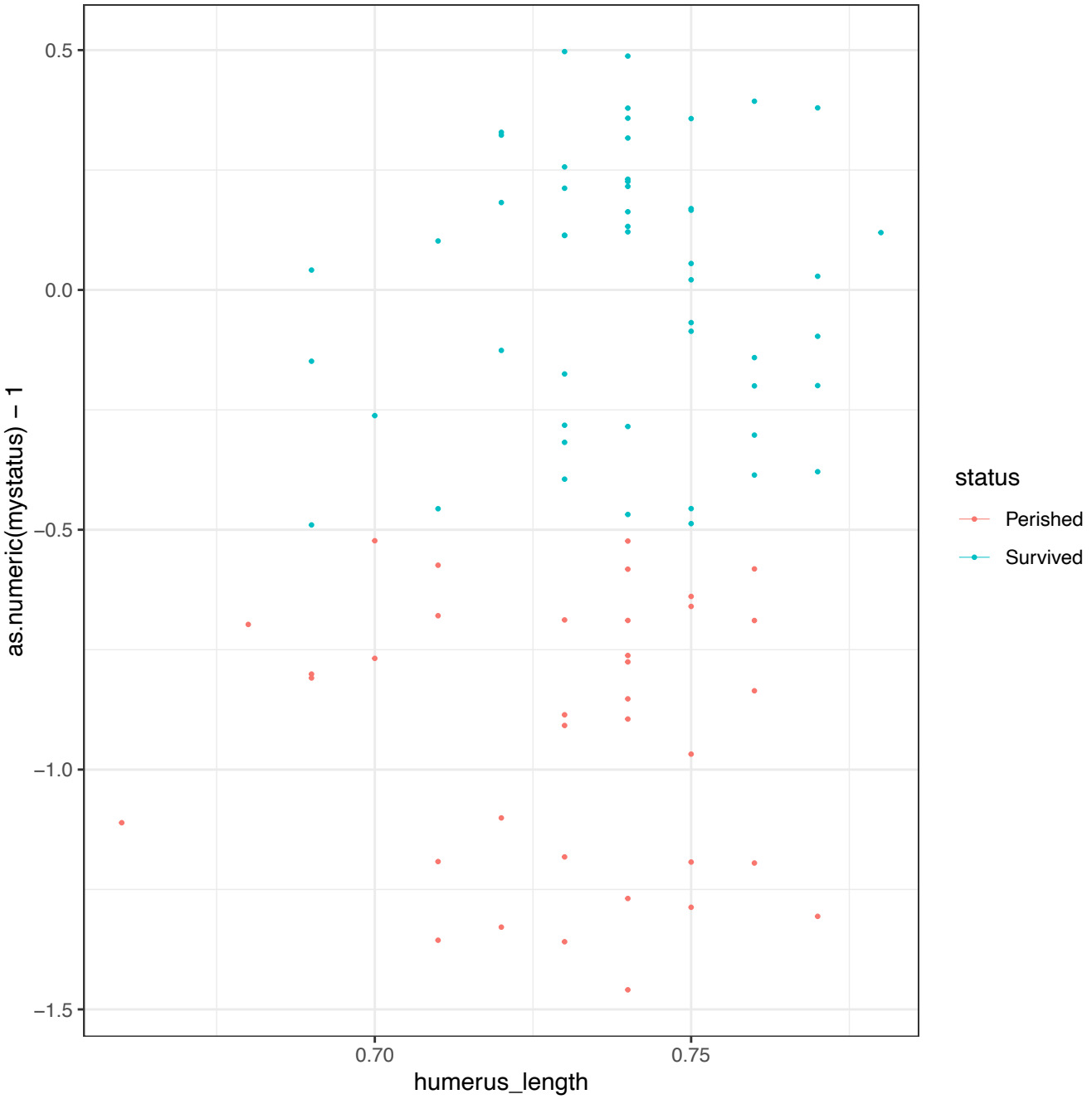
```
#It is NOT important in this case.
```

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?

```
#total_length is important and the change in odds survival from a total length increaser from 160 to 165 would
```

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

```
#Created ggplot
ggplot(bump,aes(x=humerus_length, y=as.numeric(mystatus)-1))+theme_bw()+geom_smooth(aes(color=status),method =
```



\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.

```
library(rattle)
data("wine")
view(wine)

#The wine data set includes 14 variables and 178 obs. Out of those variables,
#the first one is type which consists of three types. The rest, are 13 chemical element variables.
```

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

```
library(caret)
library(klaR)
#The . allows to put all the variables, would've been beneficial to know earlier
LINDA<-lda(Type~.,data=wine)
#Predict value based on the input data
PLINDA<-predict(LINDA,data=wine)
#Confusion Matrix using LDA
confusionMatrix(PLINDA$class,wine$Type)

Confusion Matrix and Statistics

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

Overall Statistics

               Accuracy : 1
                 95% CI : (0.9795, 1)
    No Information Rate : 0.3989
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            1.0000   1.0000   1.0000
Specificity            1.0000   1.0000   1.0000
Pos Pred Value         1.0000   1.0000   1.0000
Neg Pred Value         1.0000   1.0000   1.0000
Prevalence             0.3315   0.3989   0.2697
Detection Rate         0.3315   0.3989   0.2697
Detection Prevalence   0.3315   0.3989   0.2697
Balanced Accuracy      1.0000   1.0000   1.0000

#The classification error rate is equal to 0.
```

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

```
#Performing QDA on all variables
quda<-qda(Type~.,data=wine)
#Predict value based on the input data and creating confusion matrix
confusionMatrix(predict(quda)$class,wine$Type)
```

```
Confusion Matrix and Statistics

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

Overall Statistics

               Accuracy : 0.9944
                 95% CI : (0.9691, 0.9999)
    No Information Rate : 0.3989
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9915

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            1.0000   0.9859   1.0000
Specificity            0.9916   1.0000   1.0000
Pos Pred Value         0.9833   1.0000   1.0000
Neg Pred Value         1.0000   0.9907   1.0000
Prevalence             0.3315   0.3989   0.2697
Detection Rate         0.3315   0.3933   0.2697
Detection Prevalence   0.3371   0.3933   0.2697
Balanced Accuracy      0.9958   0.9930   1.0000

#The classification error rate is equal to almost 0, its accuracy is 0.994382
```

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

```
library(e1071)
#Creating a support vector machine of all variables in wine data
SUVM<-svm(Type~.,data=wine)
#Making confusion Matrix
confusionMatrix(predict(SUVM),wine$Type)

Confusion Matrix and Statistics

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

Overall Statistics

               Accuracy : 1
                 95% CI : (0.9795, 1)
    No Information Rate : 0.3989
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1
```

```
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            1.0000   1.0000   1.0000
Specificity            1.0000   1.0000   1.0000
Pos Pred Value         1.0000   1.0000   1.0000
Neg Pred Value         1.0000   1.0000   1.0000
Prevalence             0.3315   0.3989   0.2697
Detection Rate         0.3315   0.3989   0.2697
Detection Prevalence   0.3315   0.3989   0.2697
Balanced Accuracy      1.0000   1.0000   1.0000
```

```
#0
```

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)

```
#The classification error rate was the same for LDA,and SVM, the QDA was a little higher
#but by looking at the cross-validation, it can be cause by randomness in the data.
#All 3 methods are good. My preference would be LDA since it is the one I used the most.
```

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

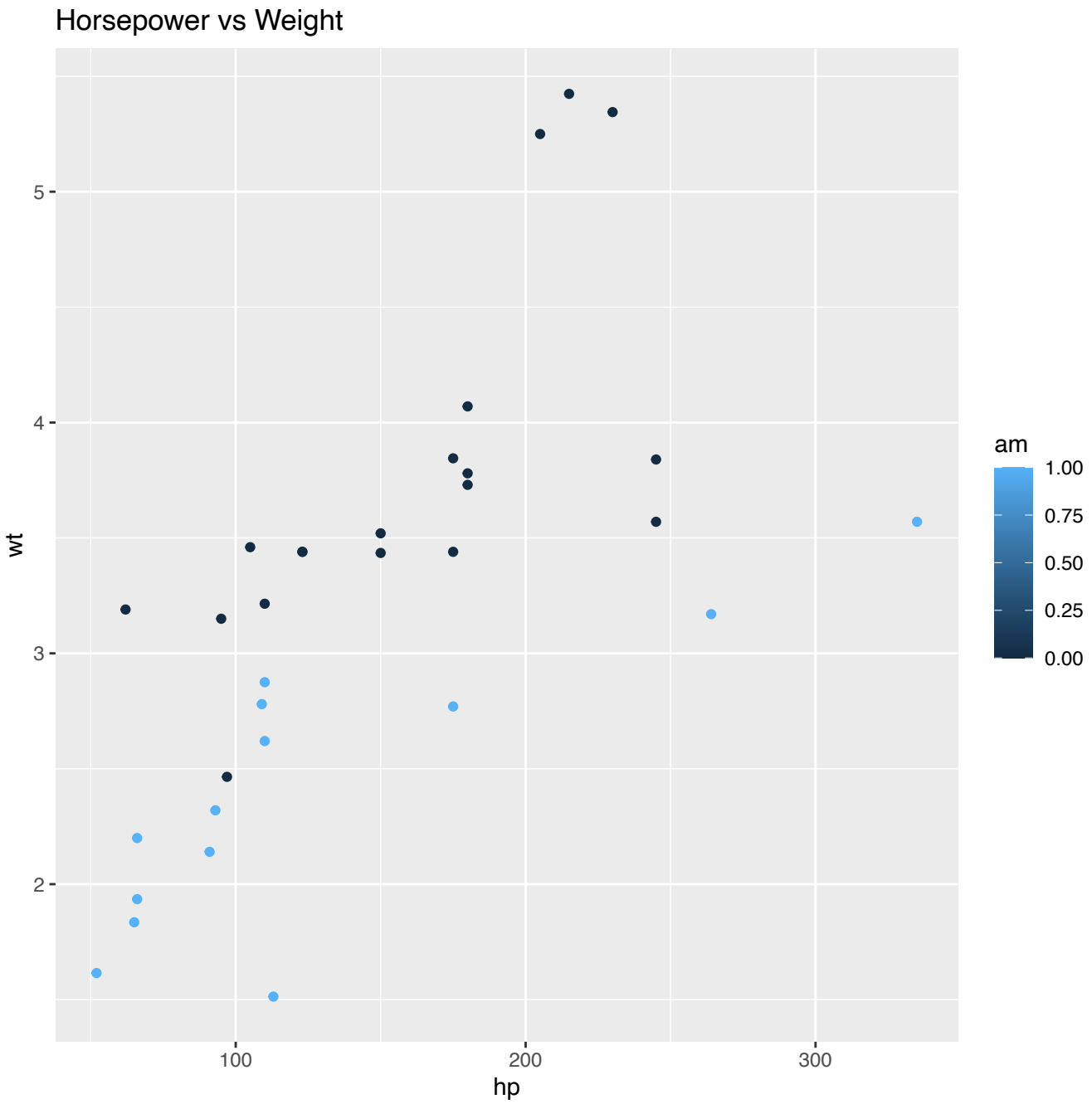
Load the `mtcars` data in R.

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

```
#am variable is described as transmission where 1 is for manual and 0 for automatic
```

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?

```
ggplot(mtcars,aes(x=hp,y=wt))+ggtitle("Horsepower vs Weight")+geom_point(aes(color=am))
```



#By looking at the graph, I can say that cars with more weight and horsepower tend to be automatic.

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?

```
carlrmwt<-glm(am~wt,data=mtcars,family=binomial)
summary(carlrmwt)
```

```
Call:
glm(formula = am ~ wt, family = binomial, data = mtcars)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.11400  -0.53738  -0.08811   0.26055   2.19931
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   12.040      4.510   2.670  0.00759 **
wt            -4.024      1.436  -2.801  0.00509 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43.230  on 31  degrees of freedom
Residual deviance: 19.176  on 30  degrees of freedom
AIC: 23.176

Number of Fisher Scoring iterations: 6

#Heavier cars are less likkely to have manual transmissions because
#if weight increases by 1000 lbs, the change in odds of a car having
#manual transmission is around 98.21% and the odds in favor of manual transmission is of 0.0178813.
```

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?

```
carlrmhp<-glm(am~hp,data=mtcars,family=binomial)
summary(carlrmhp)

Call:
glm(formula = am ~ hp, family = binomial, data = mtcars)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2955  -0.9968  -0.7818   1.1630   2.0379

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.776614   0.915429   0.848    0.396
hp          -0.008117   0.006074  -1.336    0.181

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43.230  on 31  degrees of freedom
Residual deviance: 41.228  on 30  degrees of freedom
AIC: 45.228

Number of Fisher Scoring iterations: 4

# Higer hp cars have an odds in favor of manual transmission of 0.4441024
#which means they are less likely to have manual transmissions and for every
#100 increase in hp, the change in odds of a car having manual transmission decreases by 55.59%.
```

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

```
#Because the change of odds, I would chose the weight model since it is way higher
#than the change of odds of horsepower.
```
