MATH 4322 Homework 3

Cathy Poliak

Fall 2021

Problem 1

Suppose we collect data for a group of students in a statistics class with variables X_1 = hours studied, X_2 =undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$.

(a) Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class.
(b) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?

Answer

(a) The model is:

$$p(\hat{X}) = \frac{exp(-6 + 0.05 \times \text{hours} + \text{GPA})}{1 + exp(-6 + 0.05 \times \text{hours} + \text{GPA})}$$

Thus $p(\hat{X}) = 0.3775$.

(b) Use this as the model:

$$\begin{aligned} log\left(\frac{p(X)}{1 - p(X)}\right) &= -6 + 0.05h + 3.5 \\ \log(1) &= -2.5 + 0.05h \\ 2.5 &= 0.05h \\ h &= 50 \end{aligned}$$

Problem 2

In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the `Auto` data set in the `ISLR` package.

(a) Create a binary variable, `mpg01`, that contains a 1 if mpg contains a value above its median, and a 0 if mpg contains a value below its median. You can compute the median using the `median()` function. Note you may find it helpful to use the `data.frame()` function to create a single data set containing both `mpg01` and the other Auto variables.

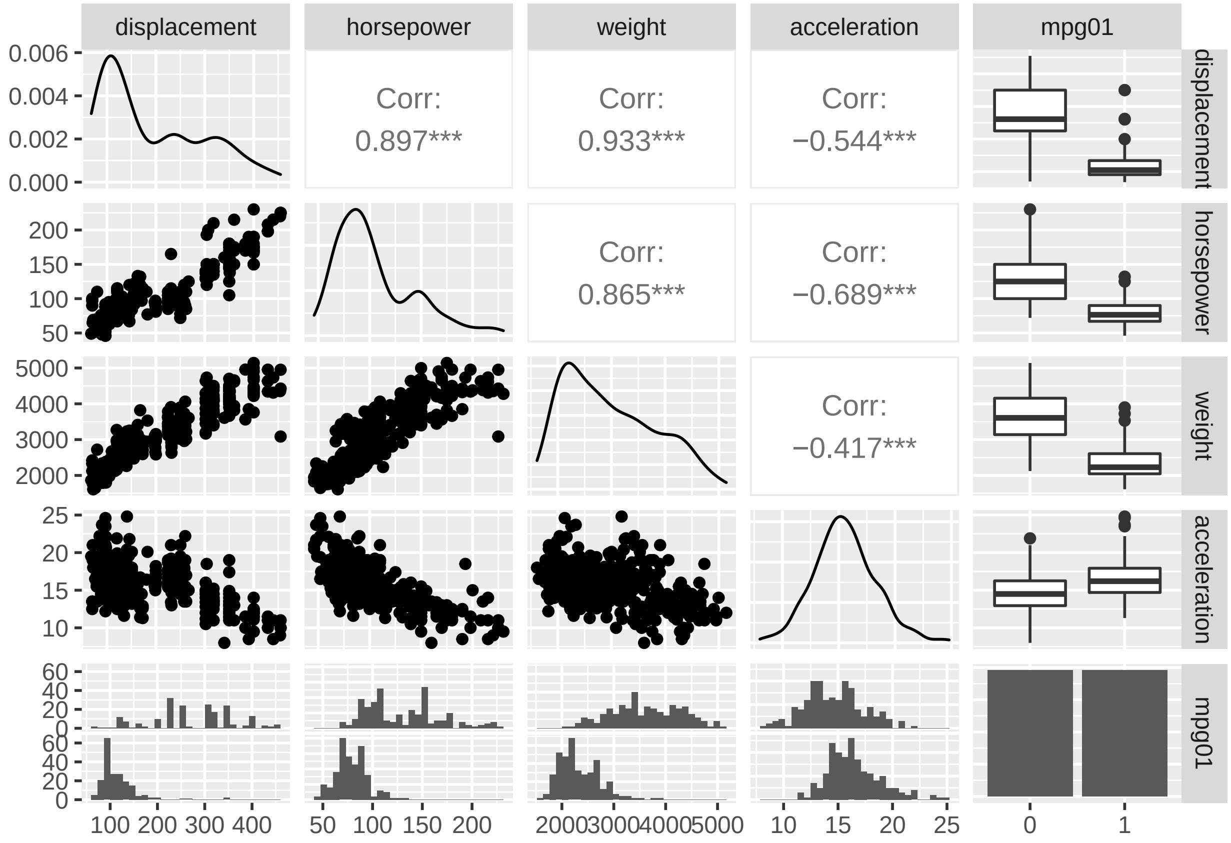
Answer

```
library(ISLR)
mpg01 = ifelse(Auto$mpg >= median(Auto$mpg),1,0)
mpg01 = factor(mpg01)
auto.new = data.frame(Auto,mpg01)
auto.new$horsepower = as.numeric(auto.new$horsepower)
```

(b) Explore the data graphically in order to investigate the association between `mpg01` and the other features. Which of the other features seem most likely to be useful in predicting `mpg01`? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

Answer

```
library(ggplot2)
library(GGally)
auto.new$cylinders = factor(auto.new$cylinders)
auto.new$year = factor(auto.new$year)
auto.new$origin = factor(auto.new$origin)
ggpairs(auto.new[,c(3:6,10)])
```



It appears that `displacement`, `horsepower` and `weight` are associated with `mpg01`.

(c) Split the data into a training set and a test set.

Answer

```r
set.seed(10)
sample = sample.int(n = nrow(auto.new),
                    size = floor(.75*nrow(auto.new)),
                    replace = F)
train = auto.new[sample,]
test = auto.new[-sample,]
```

(d) Perform logistic regression on the training data in order to predict `mpg01` using the variables that seemed most associated with `mpg01` in (b). What is the test error of the model obtained? That is use the test data to predict and get the confusion matrix and determine the error rate.

Answer

```r
auto.glm = glm(mpg01 ~  displacement + horsepower + weight, data =  train, family = "binomial")
summary(auto.glm)
```

```
##
## Call:
## glm(formula = mpg01 ~ displacement + horsepower + weight, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3739  -0.2023  -0.0052   0.4063   3.3084
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  10.3144688  1.6769444   6.151 7.71e-10 ***
## displacement -0.0156996  0.0060420  -2.598  0.00937 **
## horsepower   -0.0390595  0.0152233  -2.566  0.01029 *
## weight       -0.0013861  0.0007581  -1.828  0.06750 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 407.45  on 293  degrees of freedom
## Residual deviance: 165.79  on 290  degrees of freedom
## AIC: 173.79
##
## Number of Fisher Scoring iterations: 7
```

```r
glm.pred = predict(auto.glm, newdata = test, type = "response")
yHat = glm.pred > 0.5
table(test$mpg01,yHat)
```

```
##    yHat
##     FALSE TRUE
##   0    41    5
##   1     3   49
```

Test error rate = 8/98 = 0.0816

(e) Perform LDA on the training data in order to predict `mpg01` using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained? That is use the test data to predict and get the confusion matrix and determine the error rate.

Answer

```
library(MASS)
auto.lda = lda(mpg01 ~ displacement + horsepower + weight, data = train)
lda.pred = predict(auto.lda,test)
table(test$mpg01,lda.pred$class)
```

```
##
##       0  1
##    0 37  9
##    1  2 50
```

Error rate = 11/97 = 0.1134