Homework 3

Yiming Gao (NetID: yimingg2)

2017/2/12

Contents

Exercise 1 1

Linear Regression **1**

Exercise 2 **3**

Exercise 3 **3**

Please see the homework instructions document for detailed instructions and some grading notes. Failure to follow instructions will result in point reductions.

For this homework we return to the data found in `auto-train.csv` and `auto-test.csv` which contain train and test data respectively. `auto.csv` is provided but not used. It is a modification of the `Auto` data from the `ISLR` package.

We will use this data for each exercise in this homework.

For information on the original data:

Exercise 1

Linear Regression

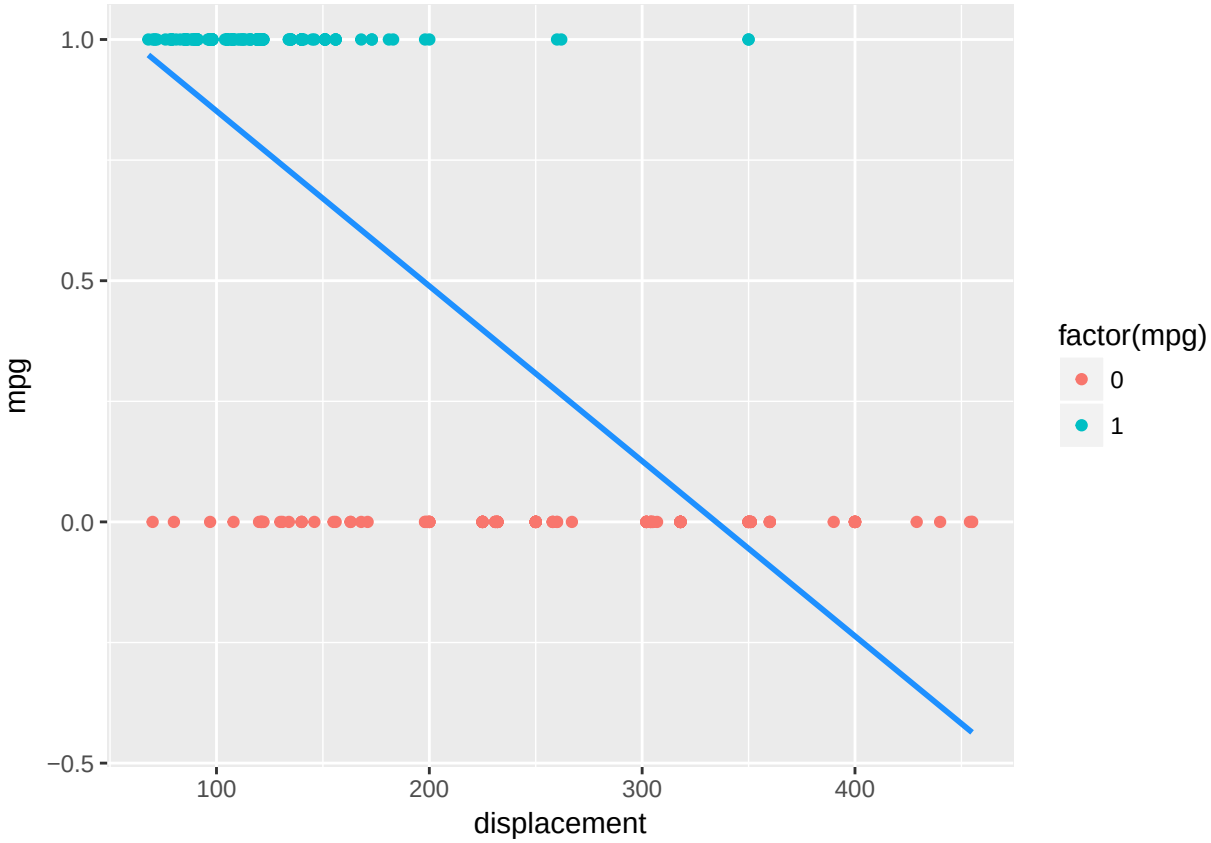
```
# create the linear model
model_lm = lm(mpg ~ displacement, data = auto_train_data)
```

Then we plot the training data and add a line with the predicted probabilities.

```
library(ggplot2)
mydata1 = data.frame("displacement" = auto_train_data$displacement, "mpg" = auto_train_data$mpg

ggplot(mydata1, aes(displacement, mpg))+
  # use colors based on the response category
```

```
  geom_point(aes(colour = factor(mpg)))+
  geom_smooth(method = "lm", se = FALSE, colour = "dodgerblue")
```



We want to find the decision boundary c such that

$$\hat{C}(\text{displacement}) = \begin{cases} 0 & \text{displacement} > c \\ 1 & \text{displacement} \leq c \end{cases}$$

Since `mpg` and `displacement` has linear relationship:

$$mpg = 1.215 - 0.0036 * displacement$$

When prediction of `mpg` is greater than 0.5, than we classify it into **one**, otherwise, we classify it into **zero**. When $mpg = 0.5$, we have $displacement = 196.9697$. So we let $c = 196$, i.e.

$$\hat{C}(\text{displacement}) = \begin{cases} 0 & \text{displacement} > 196 \\ 1 & \text{displacement} \leq 196 \end{cases}$$

- Report the test accuracy.

Exercise 2

[**12 points**] Now consider a logistic regression that considers two predictors, `acceleration` and `weight` in an additive model. Do the following:

- Plot the training data with `acceleration` as the x axis, and `weight` as the y axis, with the points colored according to their class. Add a line which represents the decision boundary for a classifier using 0.5 as a cutoff for predicted probability. **This may be challenging.**
- Report test sensitivity, test specificity, and test accuracy for three classifiers, each using a different cutoff for predicted probability:
  - 0.2
  - 0.5
  - 0.8
- Plot an ROC curve and report the AUC.

Exercise 3

[**8 points**] Finally, consider the full additive logistic regression. Create an improved model for classification by adding (or removing) complexity. Report relevant metrics for both models to justify your model.