Practical Machine Learning - Week 2

Saul Lugo

January 11, 2016

Splitting Data, Plotting Predictors and Training Models

The following are examples of how to split the data set in training and testing sets, how to train the model and how to plot the predictors to analyze the relationship between the predictors and the outcome.

Loading the Data

Is this example, the ISLR packages is used. This package has a dataset of Wages in the US.

```
require(ISLR); require(ggplot2); require(caret);
```

```
## Loading required package: ISLR
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.1.3
```

```
## Loading required package: caret
```

```
## Warning: package 'caret' was built under R version 3.1.3
```

```
## Loading required package: lattice
```

```
## Warning: package 'lattice' was built under R version 3.1.3
```

```
data(Wage)
head(Wage)
```

```
##        year age     sex           maritl     race       education
## 231655 2006  18 1. Male 1. Never Married 1. White    1. < HS Grad
## 86582  2004  24 1. Male 1. Never Married 1. White 4. College Grad
## 161300 2003  45 1. Male       2. Married 1. White 3. Some College
## 155159 2003  43 1. Male       2. Married 3. Asian 4. College Grad
## 11443  2005  50 1. Male      4. Divorced 1. White      2. HS Grad
## 376662 2008  54 1. Male       2. Married 1. White 4. College Grad
##                     region       jobclass         health health_ins
## 231655 2. Middle Atlantic  1. Industrial      1. <=Good      2. No
## 86582  2. Middle Atlantic 2. Information 2. >=Very Good      2. No
## 161300 2. Middle Atlantic  1. Industrial      1. <=Good     1. Yes
## 155159 2. Middle Atlantic 2. Information 2. >=Very Good     1. Yes
## 11443  2. Middle Atlantic 2. Information      1. <=Good     1. Yes
## 376662 2. Middle Atlantic 2. Information 2. >=Very Good     1. Yes
##         logwage      wage
```

```
## 231655 4.318063  75.04315
## 86582  4.255273  70.47602
## 161300 4.875061 130.98218
## 155159 5.041393 154.68529
## 11443  4.318063  75.04315
## 376662 4.845098 127.11574
```

```
summary(Wage)
```

```
##       year           age               sex                    maritl
##  Min.   :2003   Min.   :18.00   1. Male  :3000   1. Never Married: 648
##  1st Qu.:2004   1st Qu.:33.75   2. Female:   0   2. Married      :2074
##  Median :2006   Median :42.00                    3. Widowed      :  19
##  Mean   :2006   Mean   :42.41                    4. Divorced     : 204
##  3rd Qu.:2008   3rd Qu.:51.00                    5. Separated    :  55
##  Max.   :2009   Max.   :80.00
##
##        race                   education                     region
##  1. White:2480   1. < HS Grad      :268   2. Middle Atlantic   :3000
##  2. Black: 293   2. HS Grad        :971   1. New England       :   0
##  3. Asian: 190   3. Some College   :650   3. East North Central:   0
##  4. Other:  37   4. College Grad   :685   4. West North Central:   0
##                  5. Advanced Degree:426   5. South Atlantic    :   0
##                                           6. East South Central:   0
##                                           (Other)              :   0
##            jobclass               health      health_ins      logwage
##  1. Industrial :1544   1. <=Good     : 858   1. Yes:2083   Min.   :3.000
##  2. Information:1456   2. >=Very Good:2142   2. No : 917   1st Qu.:4.447
##                                                            Median :4.653
##                                                            Mean   :4.654
##                                                            3rd Qu.:4.857
##                                                            Max.   :5.763
##
##       wage
##  Min.   : 20.09
##  1st Qu.: 85.38
##  Median :104.92
##  Mean   :111.70
##  3rd Qu.:128.68
##  Max.   :318.34
##
```

Splitting the Data into Training and Test set

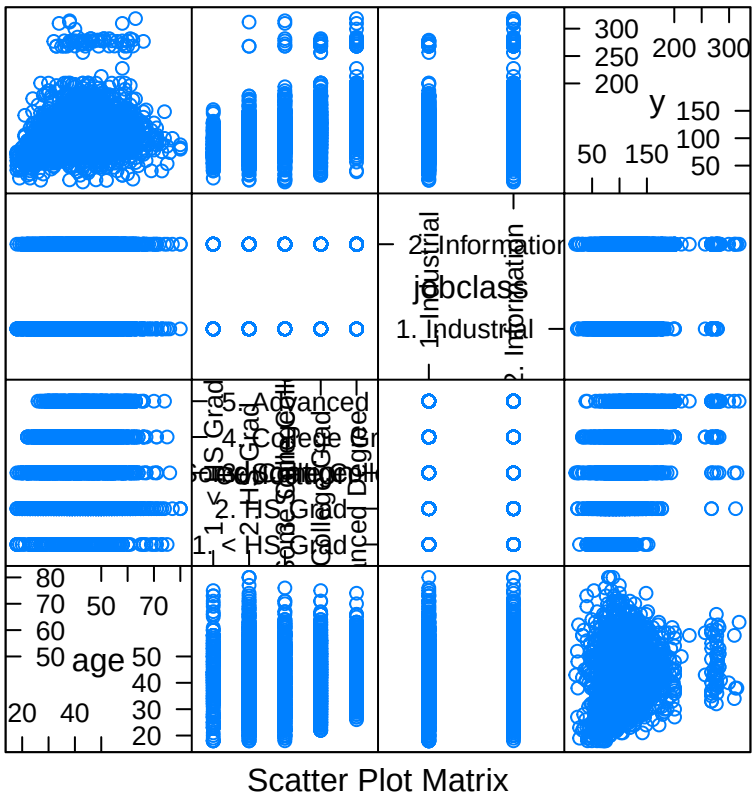
```
inTrain <- createDataPartition(y = Wage$wage, p = 0.7, list = FALSE)
training <- Wage[inTrain,]
testing <- Wage[-inTrain,]
dim(training); dim(testing)
```

```
## [1] 2102   12
```

```
## [1] 898  12
```

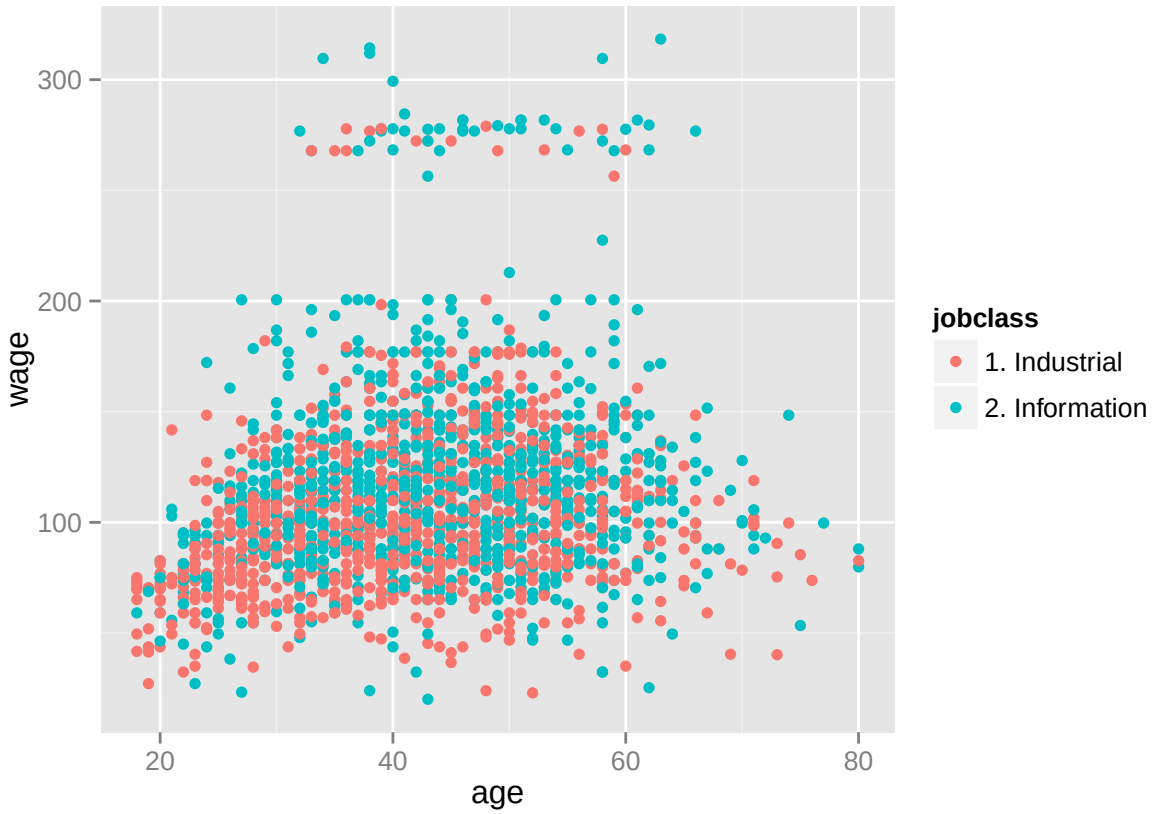
Plotting Predictors vs Outcome

```
#Plotting several predictors vs the outcome
featurePlot(x = training[,c("age", "education", "jobclass")], y = training$wage, plot="pairs")
```

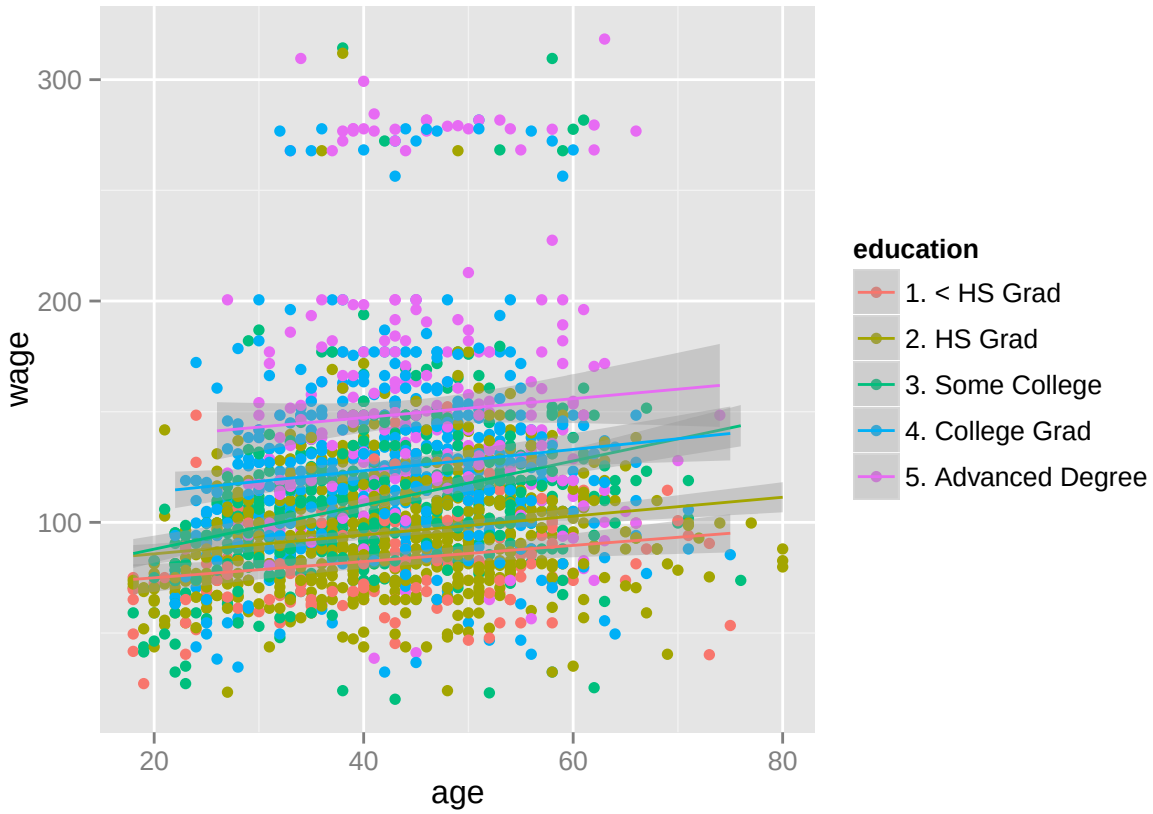


Scatter Plot Matrix

```
#Plotting one variable vs outcome and adding a second variable in the colour
qplot(age, wage, colour = jobclass,data=training)
```




```
#Add regression smoothers
qq <- qplot(age, wage, colour=education, data=training)
qq + geom_smooth(method="lm", formula = y ~ x)
```



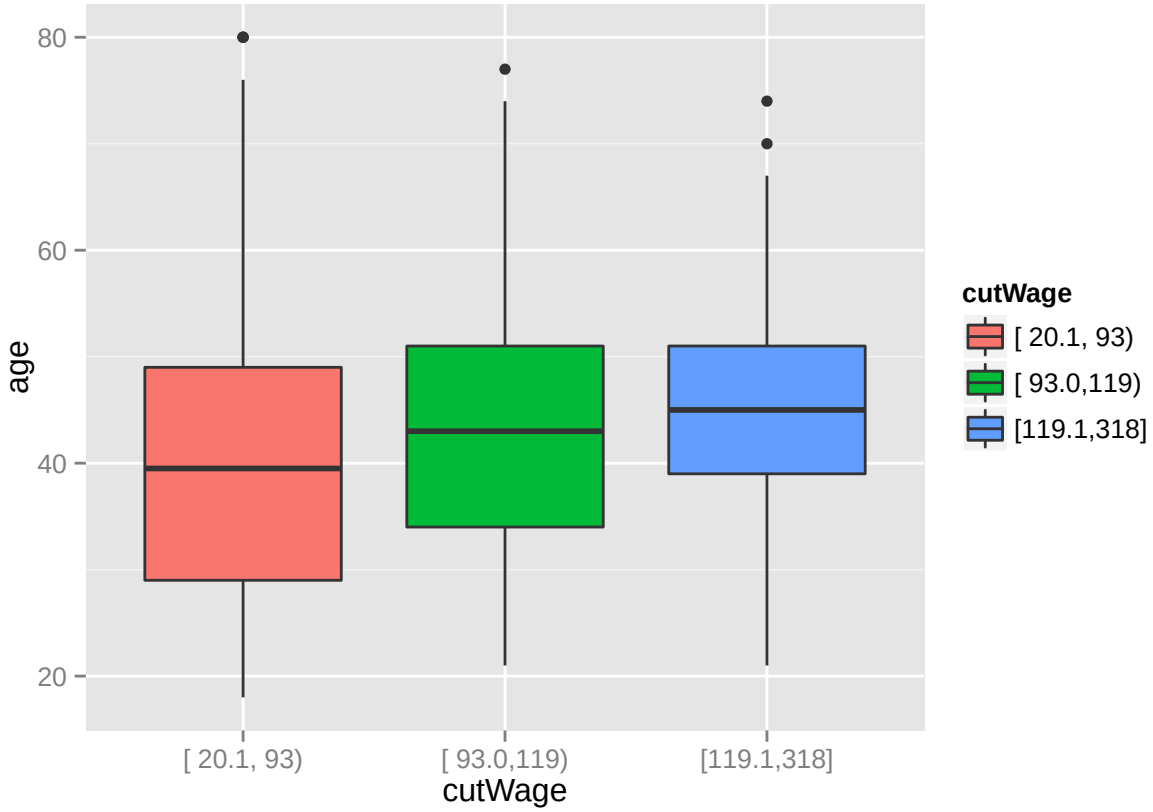

```
#cut2, making factors (Hmisc package)
require(Hmisc)
```

```
## Loading required package: Hmisc

## Warning: package 'Hmisc' was built under R version 3.1.3

## Loading required package: grid
## Loading required package: survival
##
## Attaching package: 'survival'
##
## The following object is masked from 'package:caret':
##
##     cluster
##
## Loading required package: Formula

## Warning: package 'Formula' was built under R version 3.1.3

##
## Attaching package: 'Hmisc'
##
## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

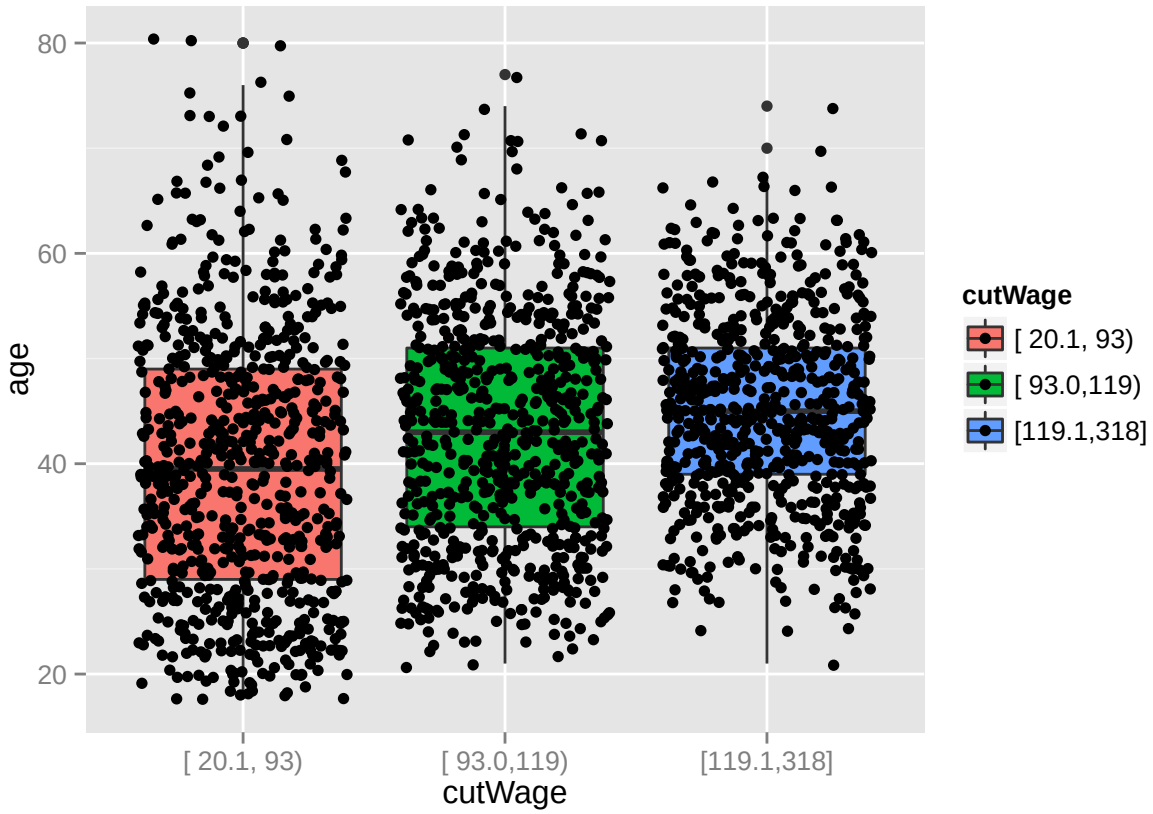
```
#Splitting the wage variable into groups of quantiles
cutWage <- cut2(training$wage, g=3)
table(cutWage)
```

```
## cutWage
## [ 20.1, 93) [ 93.0,119) [119.1,318]
##         714         711         677
```

```
#Making a boxplot to see the three different wage groups we created before
p1 <- qplot(cutWage, age, data=training, fill=cutWage, geom = c("boxplot"))
p1
```




```
#Boxplots with points overlayed
#If the jitter plot shows a lot of the points inside the boxplots that mean that the boxplots are
#actually representative of the data, so any trend one might observes might be true.
#On the contrary if only a few points are shown inside the boxplots, the trend might not be that repres
p2 <- qplot(cutWage, age, data = training, fill = cutWage, geom = c("boxplot","jitter"))
#grid.arrange(p1, p2, ncol=2)
p2
```




```
#One can make also tables
t1 <- table(cutWage,training$jobclass)
t2 <- table(cutWage,training$race)
t3 <- table(cutWage,training$education)
t1; t2; t3
```

```
##
## cutWage       1. Industrial 2. Information
##   [ 20.1, 93)           452            262
##   [ 93.0,119)           363            348
##   [119.1,318]           273            404

##
## cutWage       1. White 2. Black 3. Asian 4. Other
##   [ 20.1, 93)      569       87       43       15
##   [ 93.0,119)      594       76       33        8
##   [119.1,318]      570       50       54        3

##
## cutWage       1. < HS Grad 2. HS Grad 3. Some College 4. College Grad
##   [ 20.1, 93)          130        329             142              88
##   [ 93.0,119)           42        231             209             157
##   [119.1,318]           12         94             129             238
##
## cutWage       5. Advanced Degree
##   [ 20.1, 93)                 25
##   [ 93.0,119)                 72
##   [119.1,318]                204
```

```r
#One can also use prop.table to get the proportion on each group
prop.table(t2,1)
```

```
##
## cutWage          1. White    2. Black    3. Asian    4. Other
##   [ 20.1, 93) 0.796918768 0.121848739 0.060224090 0.021008403
##   [ 93.0,119) 0.835443038 0.106891702 0.046413502 0.011251758
##   [119.1,318] 0.841949778 0.073855244 0.079763663 0.004431315
```

```r
#Also, one can do Density Plots
qplot(wage, colour=education, data=training, geom="density")
```