Overview of "uswages" Dataset from "faraway" Package

Julian Hatwell

January 3, 2016

This document provides a brief overview of the uswages dataset in the faraway R package.

```
##       wage              educ           exper            race       
##  Min.   :  50.39   Min.   : 0.00   Min.   :-2.00   Min.   :0.000  
##  1st Qu.: 308.64   1st Qu.:12.00   1st Qu.: 8.00   1st Qu.:0.000  
##  Median : 522.32   Median :12.00   Median :15.00   Median :0.000  
##  Mean   : 608.12   Mean   :13.11   Mean   :18.41   Mean   :0.078  
##  3rd Qu.: 783.48   3rd Qu.:16.00   3rd Qu.:27.00   3rd Qu.:0.000  
##  Max.   :7716.05   Max.   :18.00   Max.   :59.00   Max.   :1.000  
##       smsa             ne              mw              so        
##  Min.   :0.000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:1.000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :1.000   Median :0.000   Median :0.0000   Median :0.0000  
##  Mean   :0.756   Mean   :0.229   Mean   :0.2485   Mean   :0.3125  
##  3rd Qu.:1.000   3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:1.0000  
##  Max.   :1.000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000  
##        we              pt        
##  Min.   :0.00   Min.   :0.0000  
##  1st Qu.:0.00   1st Qu.:0.0000  
##  Median :0.00   Median :0.0000  
##  Mean   :0.21   Mean   :0.0925  
##  3rd Qu.:0.00   3rd Qu.:0.0000  
##  Max.   :1.00   Max.   :1.0000
```

From the summary, and the associated help (not shown), the following observations can be made:

The dataframe contains 2000 rows and 10 columns.

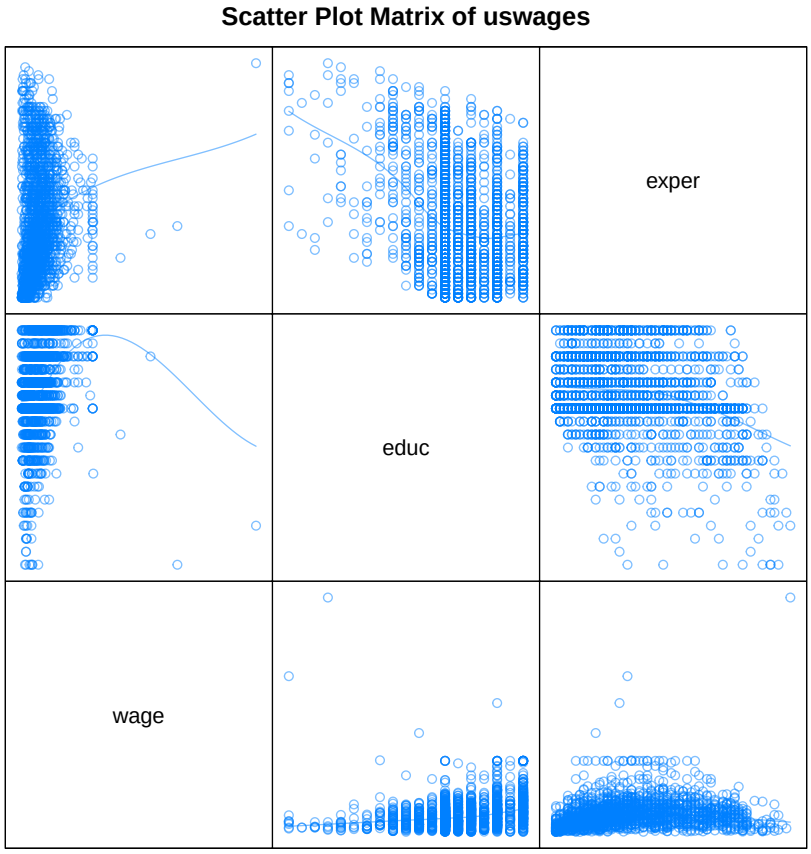

Figure 1: Correlogram

```
##
##  Welch Two Sample t-test
##
## data:  wage by race
## t = 6.1253, df = 221.05, p-value = 4.096e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  111.8771 218.0177
## sample estimates:
## mean in group white mean in group black
##            620.9838            456.0363
##
##
```

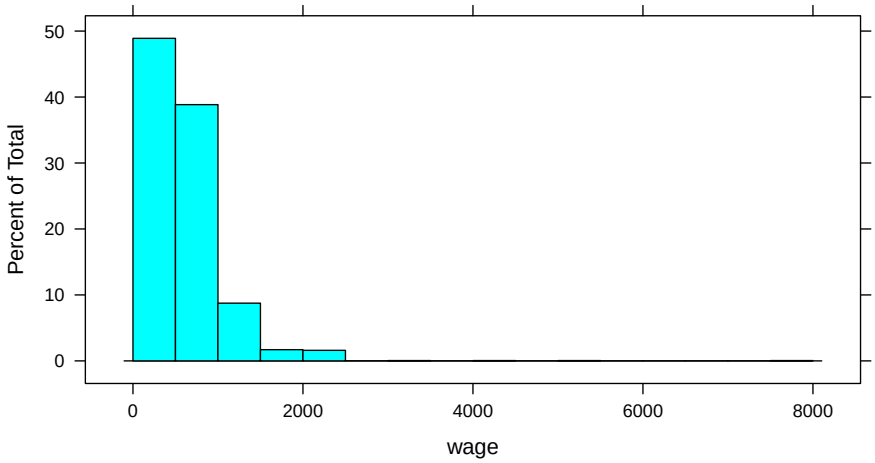

Figure 2: Histogram of the wage variable

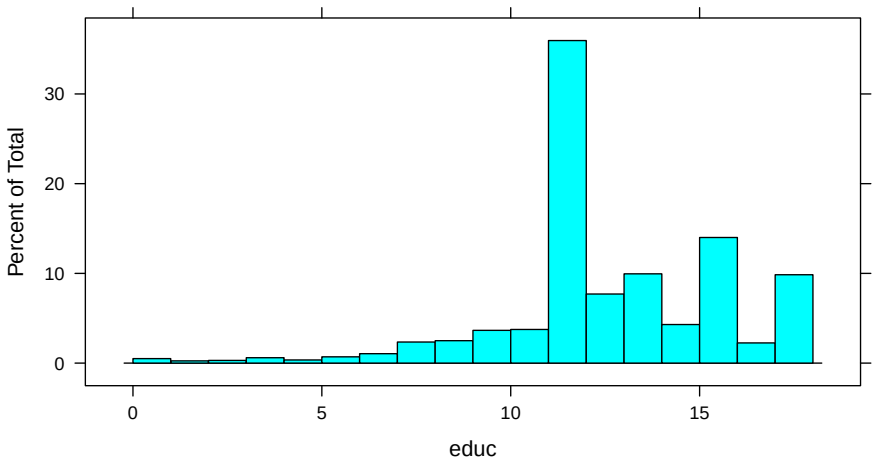

Figure 3: Histogram of the educ variable

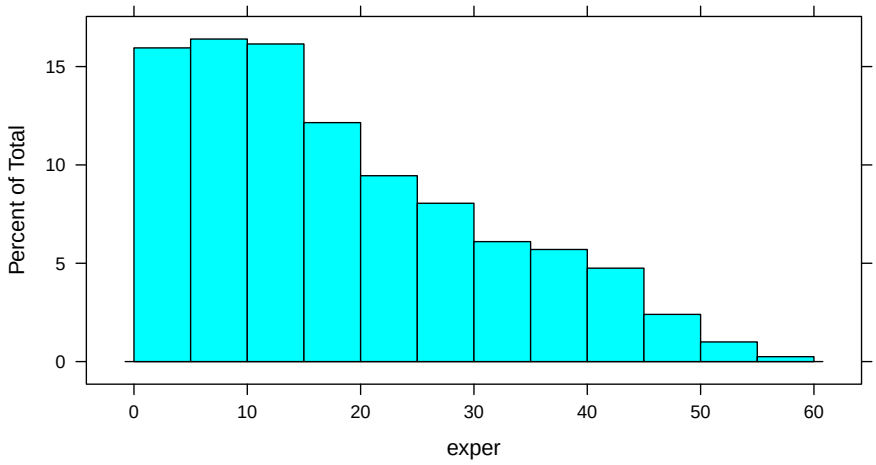

Figure 4: Histogram of the exper variable

```
##  Welch Two Sample t-test
##
## data:  wage by smsa
## t = -7.3703, df = 1185.7, p-value = 3.184e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -184.7628 -107.0756
## sample estimates:
## mean in group not in smsa     mean in group in smsa
##                   497.8030                   643.7221
##
##
##  Welch Two Sample t-test
##
## data:  wage by hours
## t = 7.4959, df = 200.4, p-value = 2.077e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  267.7346 458.8775
## sample estimates:
## mean in group full time mean in group part time
##                641.7237                278.4176
```

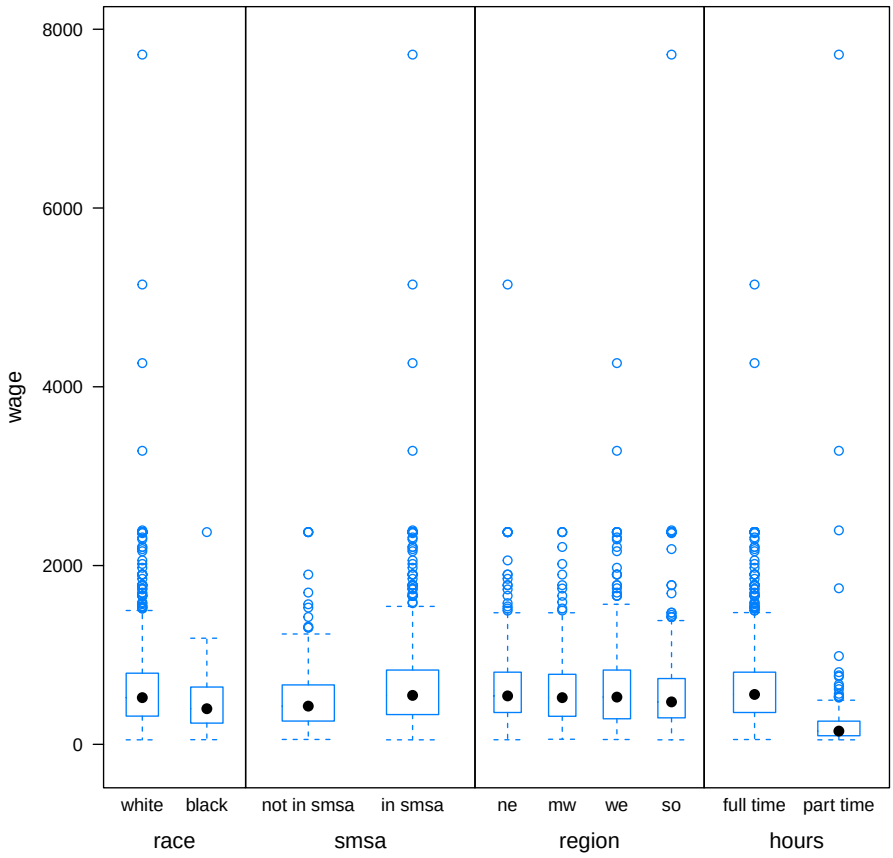

Figure 5: Boxplot of the dependent variable wage by each factor variable

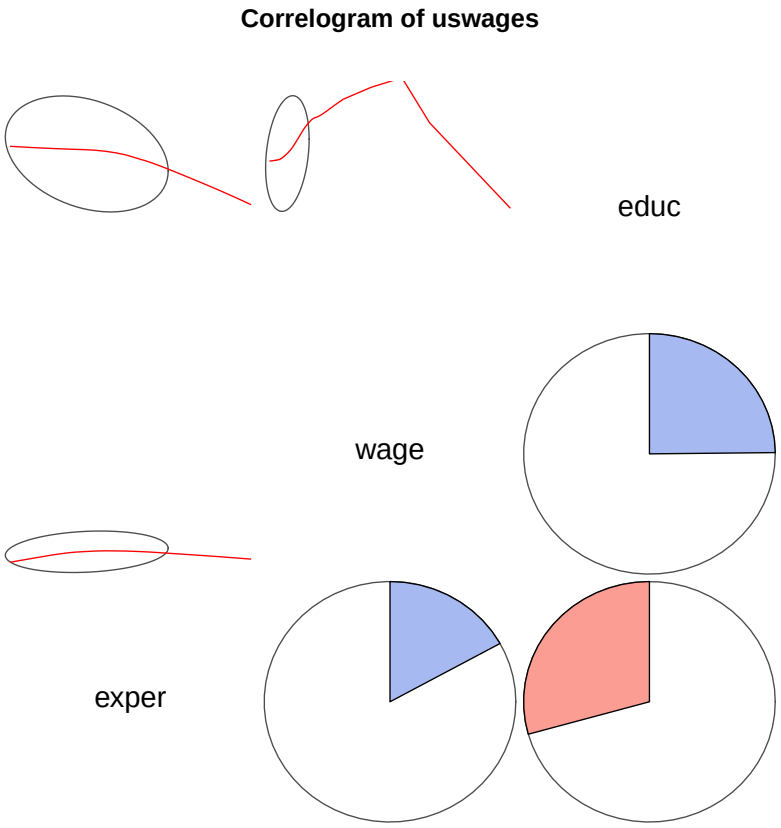

Figure 6: Correlogram

```r
# The region of origin has been coded in
# an unwieldy way and so has been
# collapsed into a single column.

# The individual with the second highest
# way has zero years education.  Zero
# years education or even low integers
# values would require verification at
# the source.  There are 40 individuals
# with educ <= 5, so it may be valid.

# 33 individuals have negative exper. Set
# these cases to NA.

# There are factor variables that don't
# add any information to the splom and
# will not be included there.

df <- df %>% gather(region, one, ne, mw,
    we, so) %>% filter(one > 0) %>% mutate(race = factor(race,
    labels = c("white", "black")), smsa = factor(smsa,
    labels = c("not in smsa", "in smsa")),
    hours = factor(pt, labels = c("full time",
        "part time")), exper = ifelse(exper <
        0, NA, exper)) %>% select(wage, educ,
    exper, race, smsa, region, hours)
```