# hw02: R Basics

*Your name*[*]

*24 10, 2018*

Questions in this problem set are taken from *Kleiber, C., & Zeileis, A. (2008). Applied econometrics with R. Springer Science & Business Media, p. 54.*

# Question 1

Create a square matrix, say $A$, with entries $a_{ii} = 2$, $i = 2, ..., n-1$, $a_{11} = a_{nn} = 1$, $a_{i,i+1} = a_{i,i-1} = -1$, and $a_{ij} = 0$ elsewhere. (Where does this matrix occur in econometrics?)

```
n = 10
diag(c(1, rep(2, n - 2), 1), n) + rbind(rep(0, n), diag(-1, n - 1, n)) + t(rbind(rep(0,
                                                                                     n), diag(-1, n - 1
```

```
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
##  [1,]    1   -1    0    0    0    0    0    0    0     0
##  [2,]   -1    2   -1    0    0    0    0    0    0     0
##  [3,]    0   -1    2   -1    0    0    0    0    0     0
##  [4,]    0    0   -1    2   -1    0    0    0    0     0
##  [5,]    0    0    0   -1    2   -1    0    0    0     0
##  [6,]    0    0    0    0   -1    2   -1    0    0     0
##  [7,]    0    0    0    0    0   -1    2   -1    0     0
##  [8,]    0    0    0    0    0    0   -1    2   -1     0
##  [9,]    0    0    0    0    0    0    0   -1    2    -1
## [10,]    0    0    0    0    0    0    0    0   -1     1
```

# Question 2

“PARADE”" is the Sunday newspaper magazine supplementing the Sunday or weekend edition of some 500 daily newspapers in the United States of America. An important yearly feature is an article providing information on some 120-150 “randomly”" selected US citizens, indicating their profession, hometown and state, and their yearly earnings. The `Parade2005` data contain the 2005 version, amended by a variable indicating celebrity status (motivated by substantial oversampling of celebrities in these data). For the `Parade2005` data:

## a)

Determine the mean earnings in California. Explain the result.

```
data("Parade2005",package="AER")
pde<-Parade2005
head(pde)
```

```
##   earnings age gender state celebrity
## 1    10000  26   male    ND        no
## 2 10000000  18 female    CA       yes
## 3    85000  39   male    NE        no
## 4    75000  50 female    NC        no
## 5    91500  61   male    DE        no
## 6    49500  39 female    SD        no
```

[*] affiliation

```
nrow(pde)
```

```
## [1] 130
```

```
cal <- subset(pde, state == "CA")
mean(cal$earnings)
```

```
## [1] 6241430
```

```
aggregate(earnings~state,pde,mean)
```

```
##    state   earnings
## 1     AK   35833.33
## 2     AL   54550.00
## 3     AZ   80000.00
## 4     CA 6241430.00
## 5     CO  156100.00
## 6     CT   80000.00
## 7     DC  120000.00
## 8     DE  185750.00
## 9     FL 6286950.00
## 10    GA   64500.00
## 11    HI  123333.33
## 12    IA   34750.00
## 13    ID   50900.00
## 14    IL   34875.00
## 15    IN 8426120.00
## 16    KS   60166.67
## 17    KY   42500.00
## 18    LA   57333.33
## 19    MA   50500.00
## 20    MD   56333.33
## 21    ME   62000.00
## 22    MI  355333.33
## 23    MN   50000.00
## 24    MO   25000.00
## 25    MS  100000.00
## 26    MT   34500.00
## 27    NC   57500.00
## 28    ND   20200.00
## 29    NE   85000.00
## 30    NH   76750.00
## 31    NJ  195250.00
## 32    NM   45000.00
## 33    NV   52700.00
## 34    NY 6033833.33
## 35    OH 5283250.00
## 36    OK   29000.00
## 37    OR   29000.00
## 38    PA   29500.00
## 39    RI   49500.00
## 40    SC   77500.00
## 41    SD   47700.00
## 42    TN   80000.00
## 43    TX 4787050.00
## 44    UT   61500.00
## 45    VA   39600.00
## 46    VT   41450.00
## 47    WA   53000.00
## 48    WI   40000.00
## 49    WV   64000.00
## 50    WY   33000.00
```

```
tapply(pde$earnings,pde$state,mean)
```

```
##         AK         AL         AZ         CA         CO         CT
##   35833.33   54550.00   80000.00 6241430.00  156100.00   80000.00
##         DC         DE         FL         GA         HI         IA
##  120000.00  185750.00 6286950.00   64500.00  123333.33   34750.00
##         ID         IL         IN         KS         KY         LA
##   50900.00   34875.00 8426120.00   60166.67   42500.00   57333.33
##         MA         MD         ME         MI         MN         MO
##   50500.00   56333.33   62000.00  355333.33   50000.00   25000.00
##         MS         MT         NC         ND         NE         NH
##  100000.00   34500.00   57500.00   20200.00   85000.00   76750.00
##         NJ         NM         NV         NY         OH         OK
##  195250.00   45000.00   52700.00 6033833.33 5283250.00   29000.00
##         OR         PA         RI         SC         SD         TN
##   29000.00   29500.00   49500.00   77500.00   47700.00   80000.00
##         TX         UT         VA         VT         WA         WI
## 4787050.00   61500.00   39600.00   41450.00   53000.00   40000.00
##         WV         WY
##   64000.00   33000.00
```

```
mean(subset(pde,state=='CA')$earnings)
```

```
## [1] 6241430
```

## b)

Determine the number of individuals residing in Idaho. (What does this say about the data set?)

```
nrow(subset(pde, state == "ID"))
```

```
## [1] 5
```

```
ida<-subset(pde,state=='ID')
ida
```

```
##    earnings age gender state celebrity
## 14    65500  36   male    ID        no
## 64    53000  40   male    ID        no
## 78    40000  43 female    ID        no
## 84    71000  42   male    ID        no
## 95    25000  36 female    ID        no
```

## c)

Determine the mean and the median earnings of celebrities. Comment.

```
options("scipen"=999)
celeb<-subset(pde,celebrity=='yes')
mean(celeb$earnings)
```

```
## [1] 17107273
```

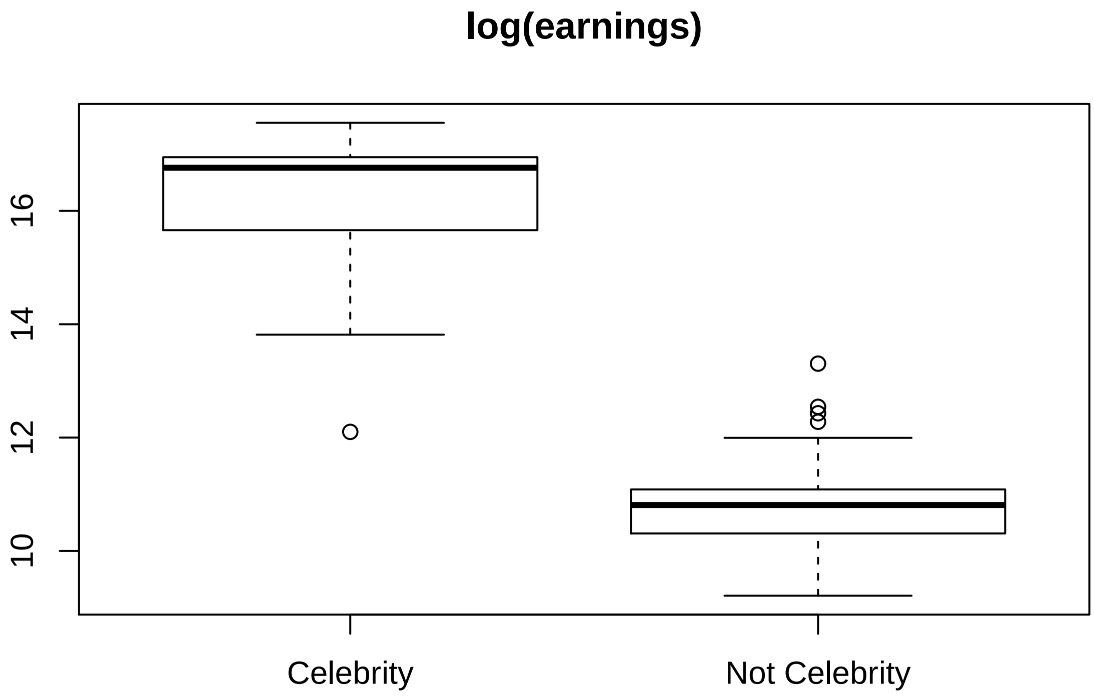
```r
median(celeb$earnings)
```

```
## [1] 19000000
```

## d)

Obtain boxplots of `log(earnings)` stratified by `celebrity`. Comment.

```r
notceleb <- subset(pde, celebrity == "no")
boxplot(log(celeb$earnings), log(notceleb$earnings), main = "log(earnings)",
        names = c("Celebrity", "Not Celebrity"))
```



# Question 3

For the `Parade2005` data of the preceding exercise, obtain a kernel density estimate of the earnings for the full data set. It will be necessary to transform the data to logarithms (why?). Comment on the result. Be sure to try out some arguments to `density()`, in particular the plug-in bandwidth `bw`.

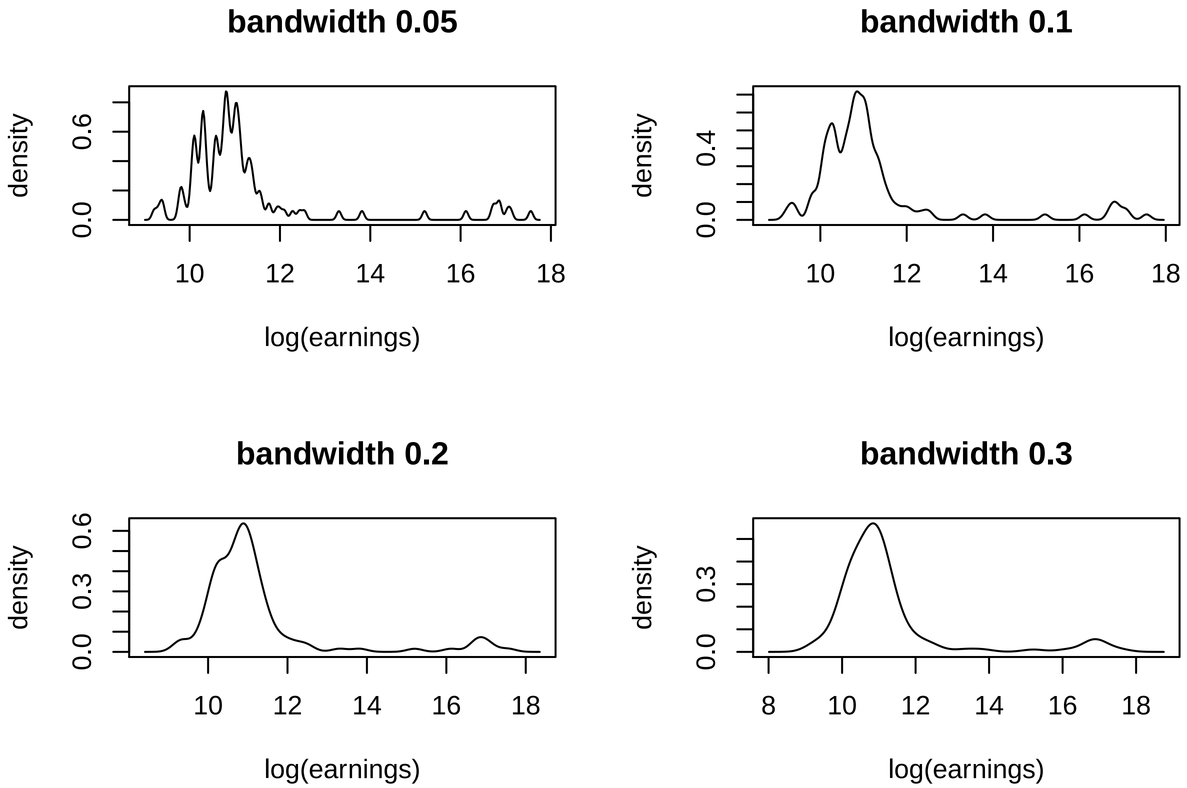
```r
library("KernSmooth")
```

```
## KernSmooth 2.23 loaded
## Copyright M. P. Wand 1997-2009
```

```r
pde<-Parade2005
summary(log(pde$earnings))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.21   10.31   10.82   11.21   11.27   17.55
```

```
pde_k05<-bkde(log(pde$earnings),bandwidth=0.05)
pde_k1<-bkde(log(pde$earnings),bandwidth=0.1)
pde_k2<-bkde(log(pde$earnings),bandwidth=0.2)
pde_k3<-bkde(log(pde$earnings),bandwidth=0.3)
par(mfrow=c(2,2))
plot(pde_k05,type="l",ylab="density",xlab="log(earnings)",main="bandwidth 0.05")
plot(pde_k1,type="l",ylab="density",xlab="log(earnings)",main="bandwidth 0.1")
plot(pde_k2,type="l",ylab="density",xlab="log(earnings)",main="bandwidth 0.2")
plot(pde_k3,type="l",ylab="density",xlab="log(earnings)",main="bandwidth 0.3")
```


bandwidth 0.05

bandwidth 0.1

bandwidth 0.2

bandwidth 0.3

```
par(mfrow=c(1,1))
```

# Question 4

Consider the `CPS1988 data`, taken from Bierens and Ginther (2001). (These data will be used for estimating an earnings equation in the next chapter.)

## a)

Obtain scatterplots of the logarithm of the real wage (`wage`) versus `experience` and versus `education`.

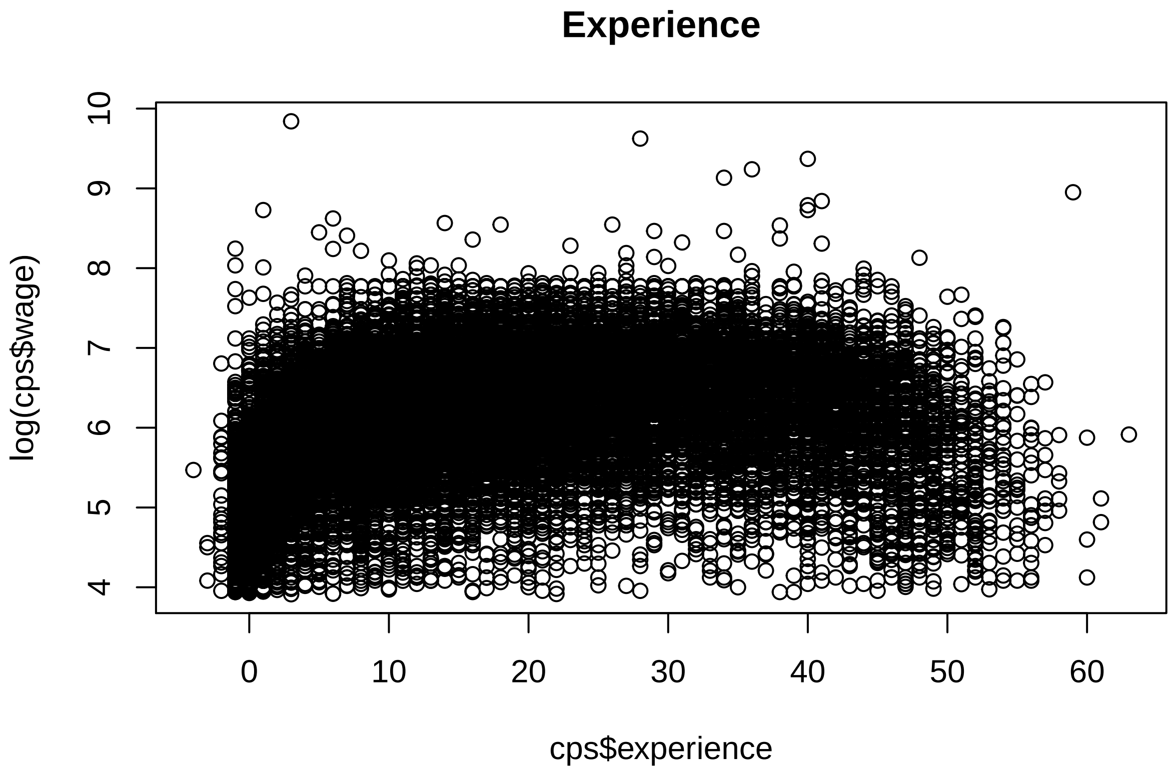
```
data("CPS1988",package="AER")
cps<-CPS1988
head(cps)
```

```
##     wage education experience ethnicity smsa    region parttime
## 1 354.94         7         45      cauc  yes northeast       no
## 2 123.46        12          1       cauc   yes northeast      yes
## 3 370.37         9          9       cauc   yes northeast       no
## 4 754.94        11         46       cauc   yes northeast       no
## 5 593.54        12         36       cauc   yes northeast       no
## 6 377.23        16         22       cauc   yes northeast       no
```

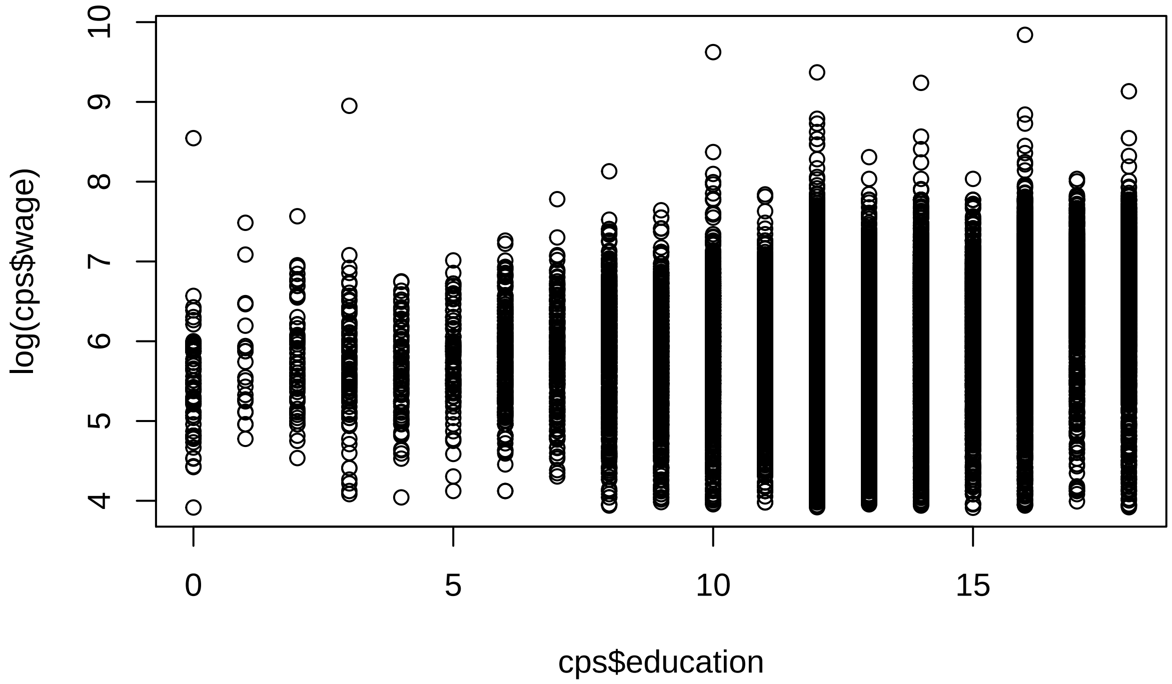
```
nrow(cps)
```

```
## [1] 28155
```

```
plot(cps$experience,log(cps$wage),main="Experience")
```



```
plot(cps$education,log(cps$wage))
```



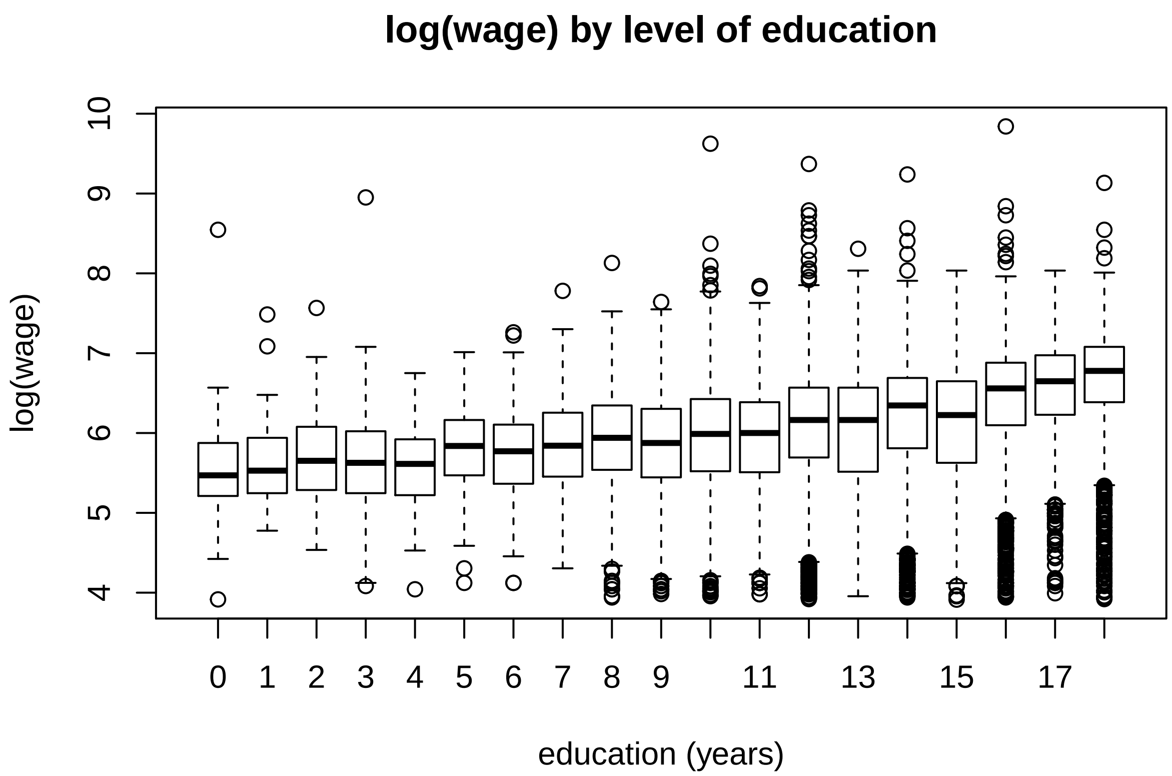
### b)

In fact, `education` corresponds to years of schooling and therefore takes on only a limited number of values. Transform `education` into a factor and obtain parallel boxplots of `wage` stratified by the levels of `education`. Repeat for `experience`.

```
summary(cps$education)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   12.00   12.00   13.07   15.00   18.00
```

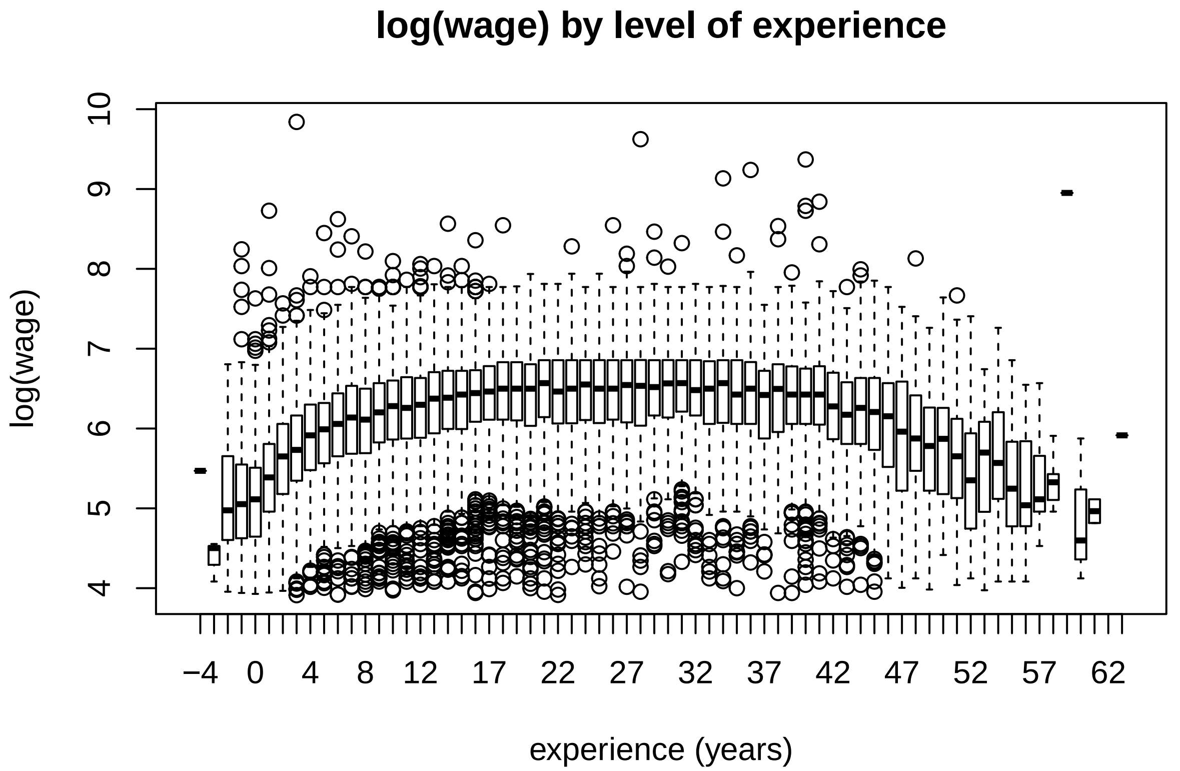
```
ed <- factor(cps$education, levels = c(rep(0:18)), labels = c(rep(0:18)))
plot(log(cps$wage) ~ ed, main = "log(wage) by level of education", xlab = "education (years)",
     ylab = "log(wage)")
```



```
summary(cps$experience)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    -4.0     8.0    16.0    18.2    27.0    63.0
```

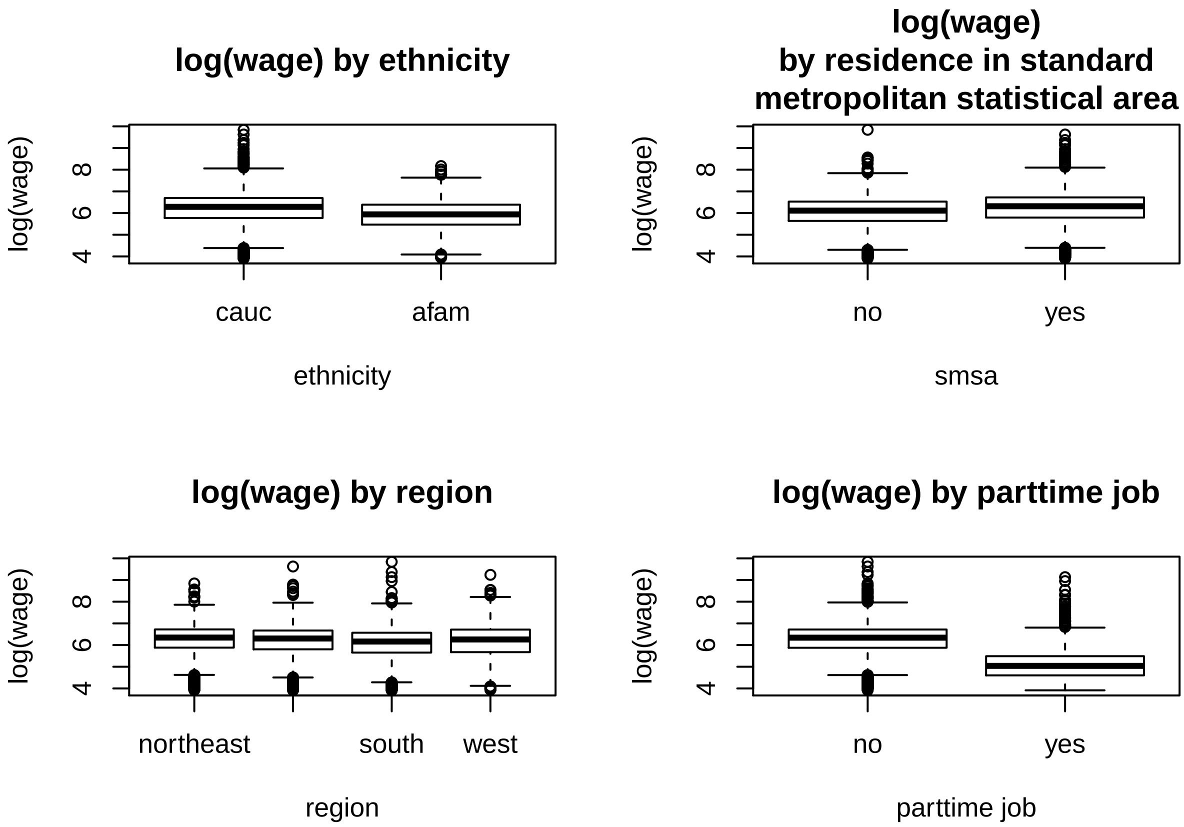
```
ex <- factor(cps$experience, levels = c(rep(-4:63)), labels = c(rep(-4:63)))
plot(log(cps$wage) ~ ex, main = "log(wage) by level of experience", xlab = "experience (years)",
     ylab = "log(wage)")
```



### c)

The data set contains four additional factors, `ethnicity`, `smsa`, `region`, and `parttime`. Obtain suitable graphical displays of log(`wage`) versus each of these factors.

```
par(mfrow = c(2, 2))
plot(log(cps$wage) ~ cps$ethnicity, main = "log(wage) by ethnicity", xlab = "ethnicity",
     ylab = "log(wage)")
plot(log(cps$wage) ~ cps$smsa, main = "log(wage)
by residence in standard
metropolitan statistical area",
     xlab = "smsa", ylab = "log(wage)")
plot(log(cps$wage) ~ cps$region, main = "log(wage) by region", xlab = "region",
     ylab = "log(wage)")
plot(log(cps$wage) ~ cps$parttime, main = "log(wage) by parttime job", xlab = "parttime job",
     ylab = "log(wage)")
```



log(wage) by ethnicity

log(wage) by residence in standard metropolitan statistical area

log(wage) by region

log(wage) by parttime job

```r
par(mfrow = c(1, 1))
```