Introductory Econometrics: Chapter 1

Author: J M Woolridge

R Code Compilation by RJ Neel

End of Chapter 1 exercises (Page 39)

Computer Exercises

C1: Use the data in WAGE1 for this exercise

```
## Computer Exercise C1
library(wooldridge) #load the Woolridge Package
wage1
?wage1 #Description of the dataset

## starting httpd help server ... done

head(wage1) #First 6 rows. Easy to view
ncol(wage1) # No of rows
nrow(wage1) #No. of columns
```

(i) Find the average education level in the sample. What are the lowest and highest years of education?

Solution

```
summary(wage1$educ)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   12.00   12.00   12.56   14.00   18.00

#Alternatively
mean(wage1$educ) #avg education level

## [1] 12.56274

min(wage1$educ) #min education leve

## [1] 0

max(wage1$educ) #max

## [1] 18
```

(ii) Find the average hourly wage in the sample. Does it seem high or low?

Solution

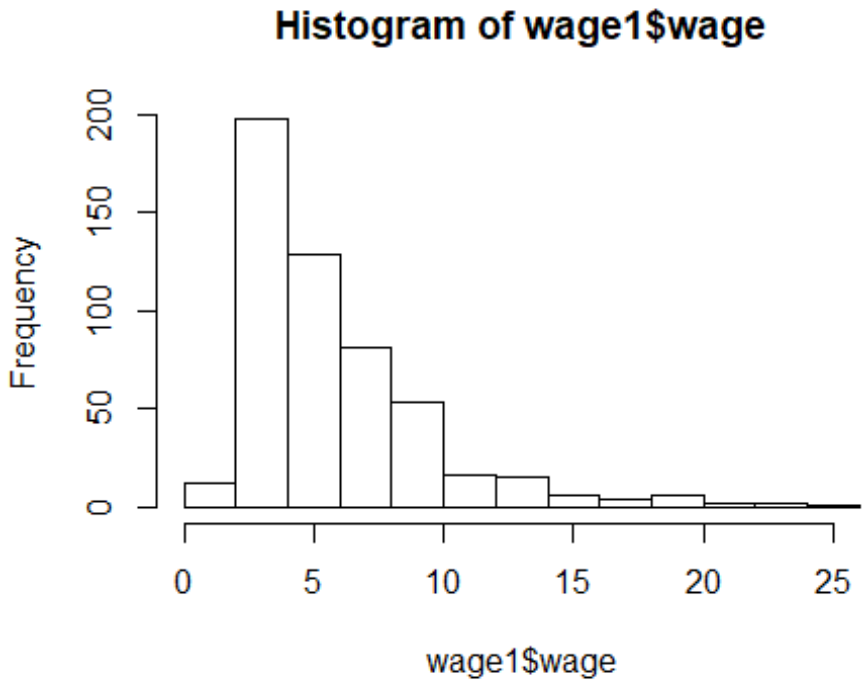
```r
mean(wage1$wage) #Gives you the average hourly wage
```

```
## [1] 5.896103
```

```r
summary(wage1$wage) #Wage appears to be low
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.530   3.330   4.650   5.896   6.880  24.980
```

```r
hist(wage1$wage) #Clearly skewed towards right
```



(iii) The wage data are reported in 1976 dollars. Using the Internet or a printed source, find the Consumer Price Index (CPI) for the years 1976 and 2013.

Solution Using Table B-60 in the 2004 Economic Report of the President, the CPI was 56.9 in 1976 and 233 in 2013.

(iv) Use the CPI values from part (iii) to find the average hourly wage in 2013 dollars. Now does the average hourly wage seem reasonable?

Solution To convert 1976 dollars into 2013 dollars, we use the ratio of the CPIs, which is $233/56.9 \approx 4.09$. Therefore, the average hourly wage in 2013 dollars is roughly $4.09(\$5.90) \approx \$24.13$, which is a reasonable figure.

(v) How many women are in the sample? How many men?

Solution

```r
head(wage1)
```

```
##   wage educ exper tenure nonwhite female married numdep smsa northcen sout
h
## 1 3.10   11     2      0        0      1       0      2    1        0
0
## 2 3.24   12    22      2        0      1       1      3    1        0
0
## 3 3.00   11     2      0        0      0       0      2    0        0
0
## 4 6.00    8    44     28        0      0       1      0    1        0
0
## 5 5.30   12     7      2        0      0       1      1    0        0
0
## 6 8.75   16     9      8        0      0       1      0    1        0
0
##   west construc ndurman trcommpu trade services profserv profocc clerocc
## 1    1        0       0        0     0        0        0       0       0
## 2    1        0       0        0     0        1        0       0       0
## 3    1        0       0        0     1        0        0       0       0
## 4    1        0       0        0     0        0        0       0       1
## 5    1        0       0        0     0        0        0       0       0
## 6    1        0       0        0     0        0        1       1       0
##   servocc    lwage expersq tenursq
## 1       0 1.131402       4       0
## 2       1 1.175573     484       4
## 3       0 1.098612       4       0
## 4       0 1.791759    1936     784
## 5       0 1.667707      49       4
## 6       0 2.169054      81      64
```

```r
#Notice the female column is a binary variable implying 1 for female and 0 fo
r male requring us to proceed with 'dplyr'
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
w=nrow(wage1 %>% group_by(female) %>% filter(female=='1'))
w
```

```
## [1] 252
```

```
m=nrow(wage1)-w
m

## [1] 274
```

End of Computer Exercise 1

C2: Use the data in BWGHT to answer this question

(i) How many women are in the sample, and how many report smoking during pregnancy?

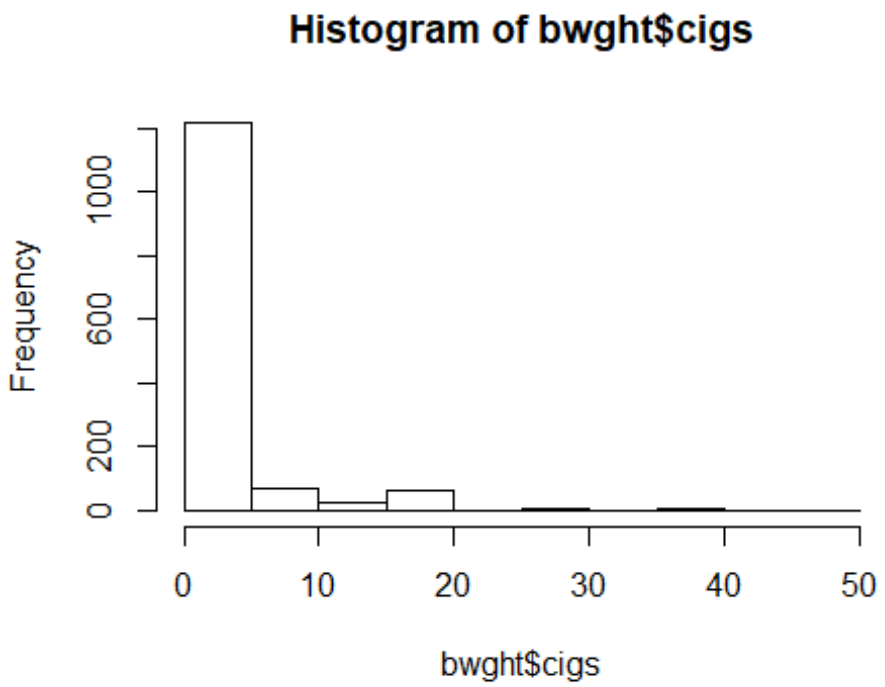
Solution

```
#Note: This data set contains all women
nrow(bwght) #No of women smoking

## [1] 1388
```

(ii) What is the average number of cigarettes smoked per day? Is the average a good measure of the "typical" woman in this case? Explain.

Solution

```
mean(bwght$cigs)

## [1] 2.087176

summary(bwght$cigs)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   2.087   0.000  50.000

hist(bwght$cigs)
```




```
#Based on the histogram and range it appears to be a good measure.
```

(iii) Among women who smoked during pregnancy, what is the average number of cigarettes smoked per day? How does this compare with your answer from part (ii), and why?

```
avg_all=mean(bwght$cigs)
avg_all
```

```
## [1] 2.087176
```

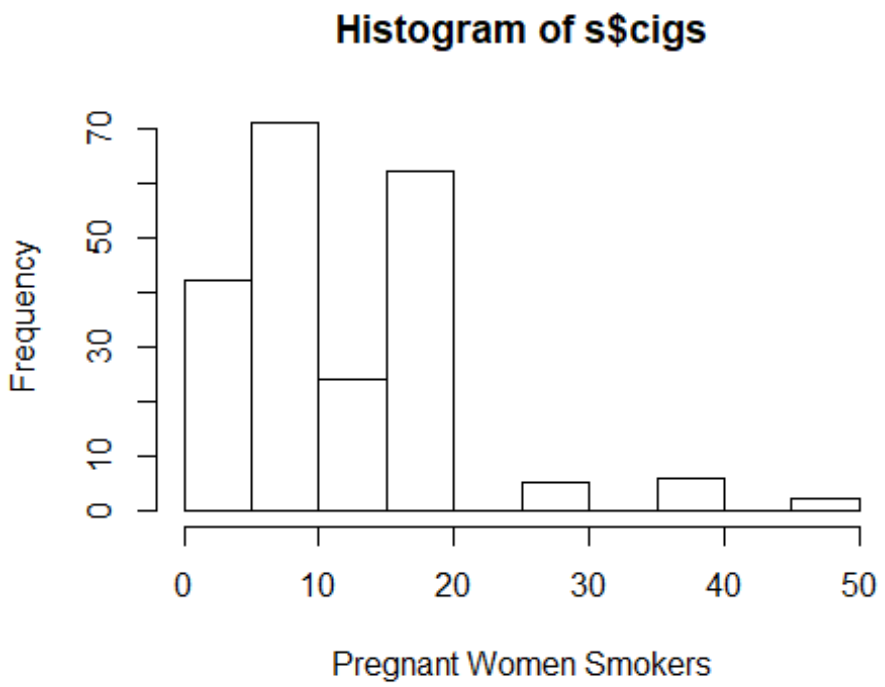
```
library(dplyr)
nrow(bwght %>% group_by(cigs) %>% filter(cigs=='0'))
```

```
## [1] 1176
```

```
s=(bwght %>% group_by(cigs) %>% filter(cigs>'0'))
avg_s=mean(s$cigs)
avg_s
```

```
## [1] 13.66509
```

```
hist(s$cigs, xlab='Pregnant Women Smokers')
```




```
#It markedly differs from the privious average by about 11 units.
```

C3 The data in MEAP01 are for the state of Michigan in the year 2001. Use these data to answer the following questions.

```
head(meap01)
```

```
##   dcode bcode math4 read4 lunch enroll  expend    exppp  lenroll  lexpend
## 1  1010  4937  83.3  77.8 40.60    468 2747475 5870.673 6.148468 14.82619
## 2  2070   597  90.3  82.3 27.10    679 1505772 2217.632 6.520621 14.22482
## 3  2080  4860  61.9  71.4 41.75    400 2121871 5304.678 5.991465 14.56781
## 4  3010   790  85.7  60.0 12.75    251 1211034 4824.836 5.525453 14.00698
## 5  3010  1403  77.3  59.1 17.08    439 1913501 4358.772 6.084499 14.46445
## 6  3010  4056  85.2  67.0 23.17    561 2637483 4701.396 6.329721 14.78534
##     lexppp
## 1 8.677725
## 2 7.704195
## 3 8.576344
## 4 8.481532
## 5 8.379946
## 6 8.455615
```

(i) Find the largest and smallest values of math4. Does the range make sense? Explain.

Solution

```
head(meap01)
```

```
##   dcode bcode math4 read4 lunch enroll  expend    exppp  lenroll  lexpend
## 1  1010  4937  83.3  77.8 40.60    468 2747475 5870.673 6.148468 14.82619
## 2  2070   597  90.3  82.3 27.10    679 1505772 2217.632 6.520621 14.22482
## 3  2080  4860  61.9  71.4 41.75    400 2121871 5304.678 5.991465 14.56781
## 4  3010   790  85.7  60.0 12.75    251 1211034 4824.836 5.525453 14.00698
## 5  3010  1403  77.3  59.1 17.08    439 1913501 4358.772 6.084499 14.46445
## 6  3010  4056  85.2  67.0 23.17    561 2637483 4701.396 6.329721 14.78534
##     lexppp
## 1 8.677725
## 2 7.704195
## 3 8.576344
## 4 8.481532
## 5 8.379946
## 6 8.455615
```

```
summary(meap01$math4) # It makes sense as a percentage is between 0 and 100
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   61.60   76.40   71.91   87.00  100.00
```

(ii) How many schools have a perfect pass rate on the math test? What percentage is this of the total sample?

Solution

```
library(dplyr)
passrate100=nrow(meap01 %>% group_by(math4) %>% filter(math4=='100'))
passrate100
```

```
## [1] 38
```

```
samplesize=nrow(meap01)
samplesize
```

```
## [1] 1823
```

```
percent_passrate=round((passrate100/samplesize)*100,2)
percent_passrate
```

```
## [1] 2.08
```

(iii) How many schools have math pass rates of exactly 50%?

Solution

```
library(dplyr)

nrow(meap01 %>% group_by(math4) %>% filter(math4=='50'))
```

```
## [1] 17
```

(iv) Compare the average pass rates for the math and reading scores. Which test is harder to pass?

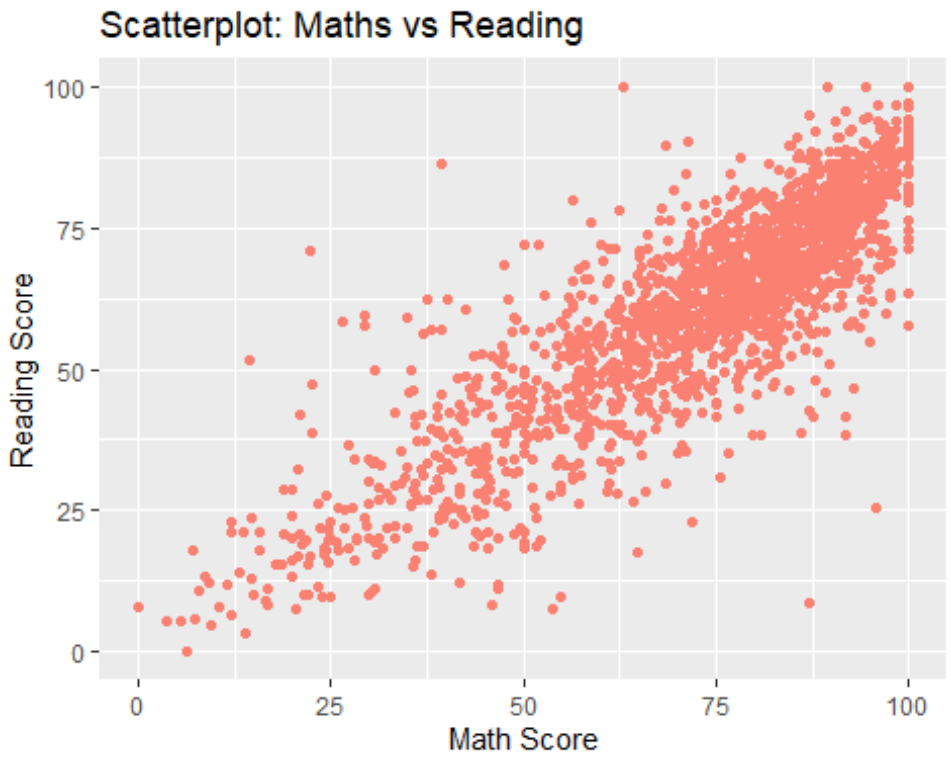
Solution

```
pass_m=mean(meap01$math4)
pass_m

## [1] 71.909

pass_r=mean(meap01$read4)
pass_r

## [1] 60.06188

#Cleary Reading is much more difficult to pass
```

(v) Find the correlation between math4 and read4. What do you conclude?

Solution

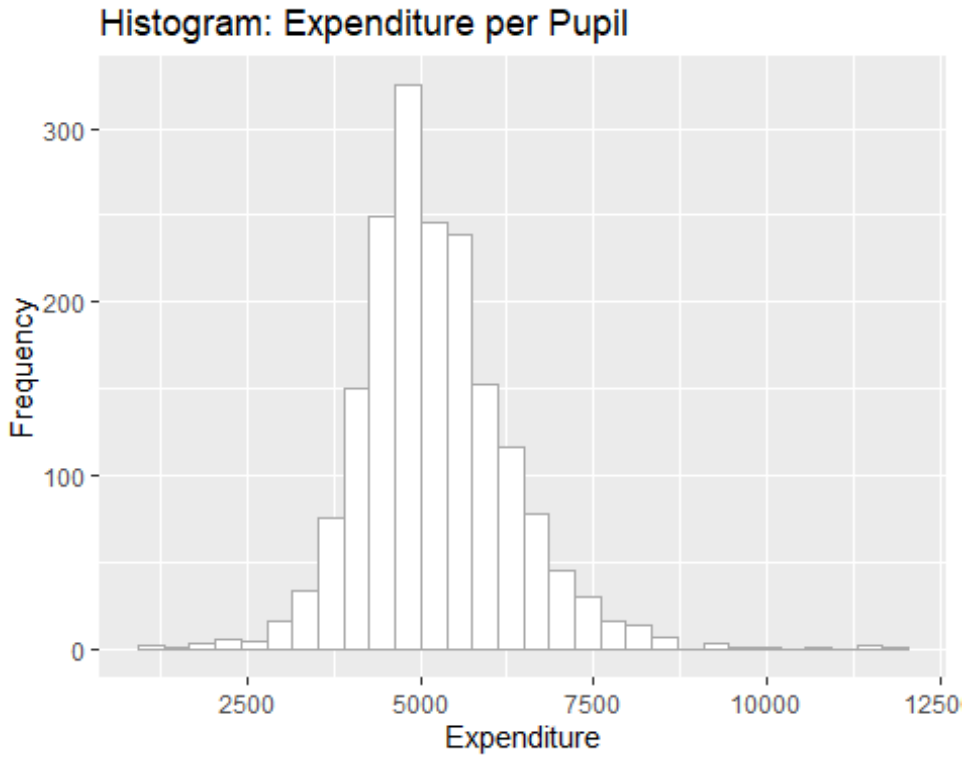
```
cor(meap01$math4,meap01$read4)

## [1] 0.8427281

library(ggplot2)
ggplot(data=meap01,aes(x=meap01$math4,y=meap01$read4))+geom_point(col='salmon
')+ggtitle("Scatterplot: Maths vs Reading")+xlab("Math Score")+ylab("Reading
Score")
```




```
#It is strongly positive
```

(vi) The variable exppp is expenditure per pupil. Find the average of exppp along with its standard deviation. Would you say there is wide variation in per pupil spending?

```
mean(meap01$exppp)

## [1] 5194.865

sd(meap01$exppp)

## [1] 1091.89

summary(meap01$exppp)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1207    4502    5078    5195    5767   11958

library(ggplot2)
ggplot(data=meap01,aes(x=meap01$exppp))+geom_histogram(col='dark grey',fill='
white')+ggtitle("Histogram: Expenditure per Pupil")+xlab("Expenditure")+ylab(
"Frequency")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
#Considering Min=1207 and Max=11958, It is significantly wide
```

(vii) Suppose School A spends $6,000 per student and School B spends $5,500 per student. By what percentage does School A's spending exceed School B's? Compare

this to 100 · [log(6,000) – log(5,500)], which is the approximation percentage difference based on the difference in the natural logs. (See Section A.4 in Appendix A.)

Solution

```
round((log(6000)-log(5500))*100,2) # Gives the Percentage

## [1] 8.7

round(((6000-5500)/ 5500)*100,2)

## [1] 9.09
```