Faraway - Chapter 1

J. A. Kilgallen

10/21/2020

Linear Models - Chapter 1 Exercises

Question 1

The following table represents a summary of data collected to study teenage gambling in Britain.

```
library(faraway)
```

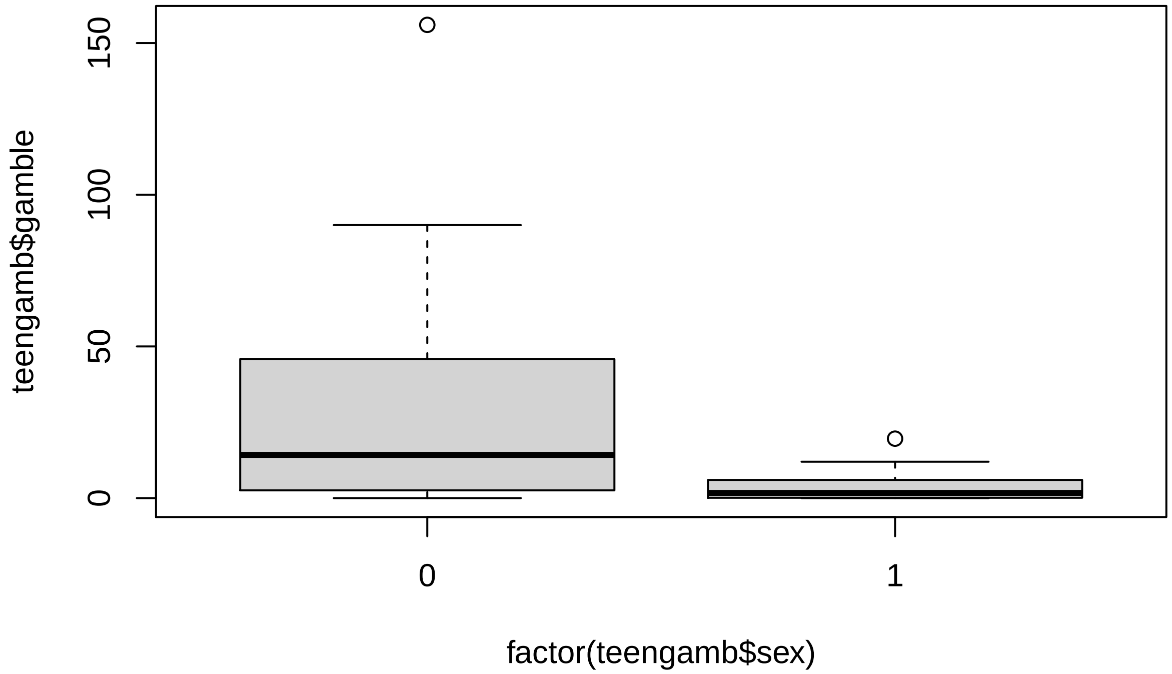
```
## Warning: package 'faraway' was built under R version 4.0.3
```

```
data(teengamb)
summary(teengamb)
```

```
##       sex             status          income           verbal     
##  Min.   :0.0000   Min.   :18.00   Min.   : 0.600   Min.   : 1.00  
##  1st Qu.:0.0000   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 6.00  
##  Median :0.0000   Median :43.00   Median : 3.250   Median : 7.00  
##  Mean   :0.4043   Mean   :45.23   Mean   : 4.642   Mean   : 6.66  
##  3rd Qu.:1.0000   3rd Qu.:61.50   3rd Qu.: 6.210   3rd Qu.: 8.00  
##  Max.   :1.0000   Max.   :75.00   Max.   :15.000   Max.   :10.00  
##      gamble     
##  Min.   :  0.0  
##  1st Qu.:  1.1  
##  Median :  6.0  
##  Mean   : 19.3  
##  3rd Qu.: 19.4  
##  Max.   :156.0  
```

We can see from this table that there were approximately 2 females for every three males in the study, and that the median amount of money spent gambling was £6. We can also give a graphical summary of some aspects of the data.In the following plot the x axis represents gender (0 for male, 1 for female), and the y axis shows the amount of money spent on gambling.

```
plot(teengamb$gamble ~ factor(teengamb$sex))
```



It can be easily observed that males on average spend much more on gambling then females. One male individual even spent £150.

Question 2

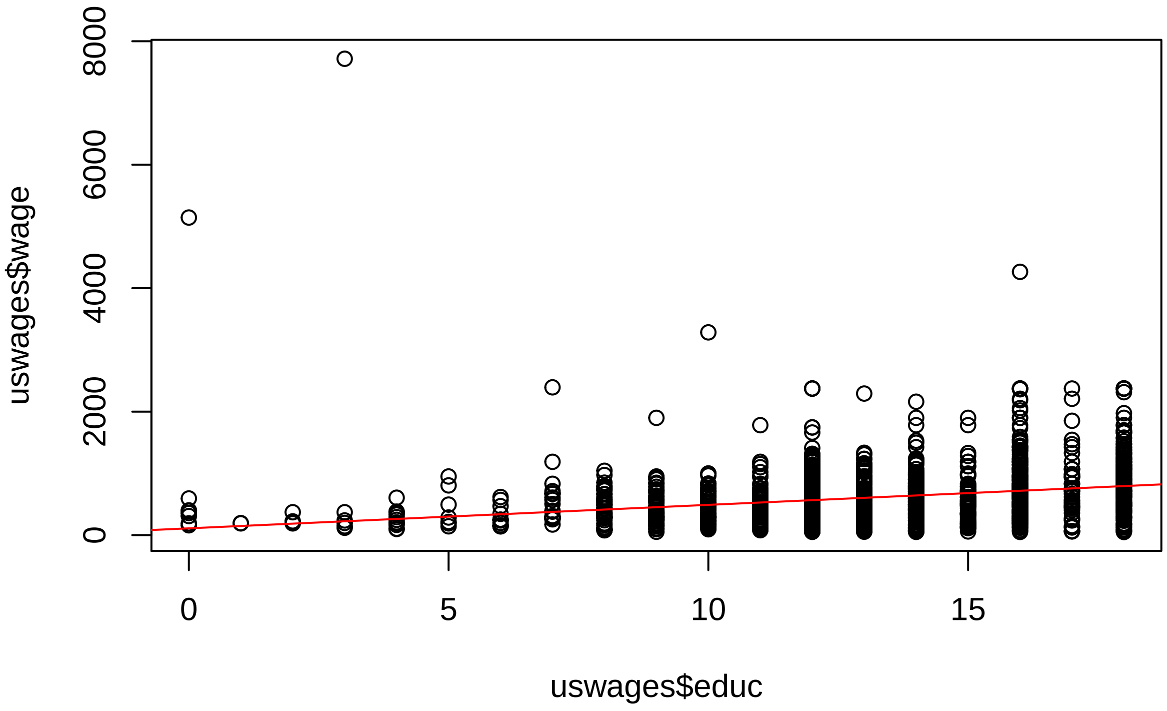
The following table represents a summary of data collected to analyse the distribution of wages in the US in 1988.

```
data("uswages")
summary(uswages)
```

```
##       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
```

We can see from the table that the mean wage was $608.12, and that there were only two categories given for race. We may also summarise some aspects of this data graphically. The following plot shows the relationship between education and wage. The x axis represents the number of years of education the person recieved, and the y axis the persons wage in dollars.

```
plot(uswages$wage ~ uswages$educ)
abline(lm(uswages$wage ~ uswages$educ), col="red")
```



We can see hear that there is a definite positive correlation between how educated an individual is and their wage.

Question 3

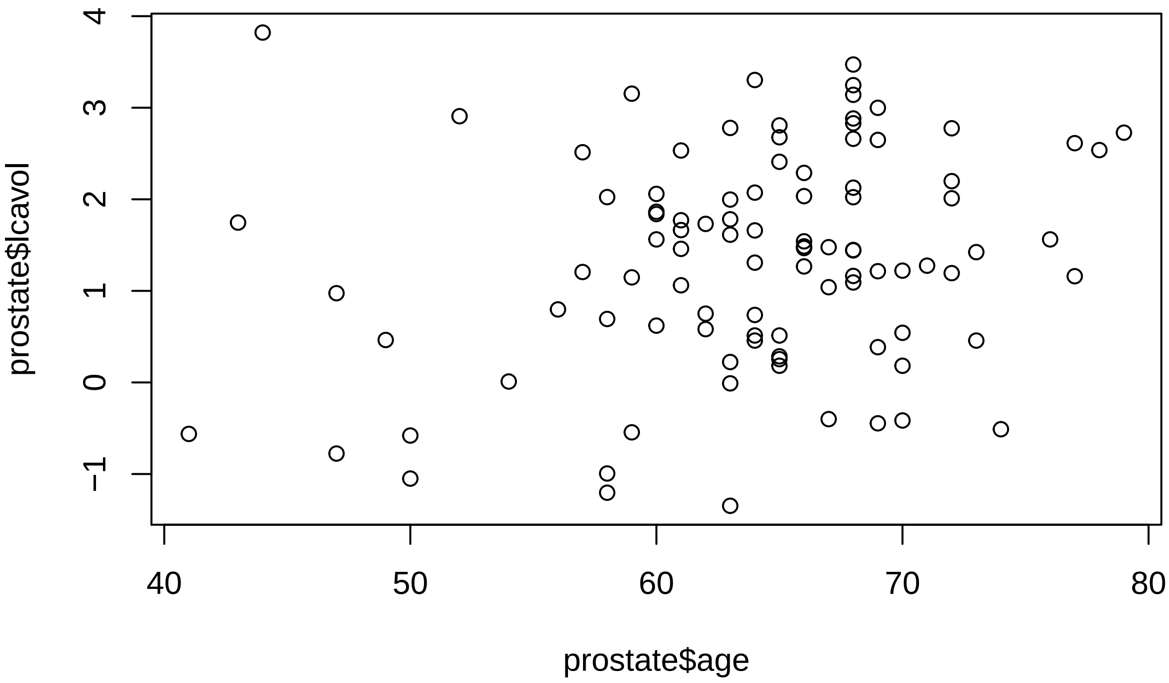
The following table represents a summary of data collected to study prostate cancer patients due to receive a radical prostatectomy.

```
data("prostate")
summary(prostate)
```

```
##      lcavol           lweight           age             lbph        
##  Min.   :-1.3471   Min.   :2.375   Min.   :41.00   Min.   :-1.3863  
##  1st Qu.: 0.5128   1st Qu.:3.376   1st Qu.:60.00   1st Qu.:-1.3863  
##  Median : 1.4469   Median :3.623   Median :65.00   Median : 0.3001  
##  Mean   : 1.3500   Mean   :3.653   Mean   :63.87   Mean   : 0.1004  
##  3rd Qu.: 2.1270   3rd Qu.:3.878   3rd Qu.:68.00   3rd Qu.: 1.5581
##  Max.   : 3.8210   Max.   :6.108   Max.   :79.00   Max.   : 2.3263
##       svi              lcp             gleason          pgg45
##  Min.   :0.0000   Min.   :-1.3863   Min.   :6.000   Min.   :  0.00
##  1st Qu.:0.0000   1st Qu.:-1.3863   1st Qu.:6.000   1st Qu.:  0.00
##  Median :0.0000   Median :-0.7985   Median :7.000   Median : 15.00
##  Mean   :0.2165   Mean   :-0.1794   Mean   :6.753   Mean   : 24.38
##  3rd Qu.:0.0000   3rd Qu.: 1.1786   3rd Qu.:7.000   3rd Qu.: 40.00
##  Max.   :1.0000   Max.   : 2.9042   Max.   :9.000   Max.   :100.00
##       lpsa
##  Min.   :-0.4308
##  1st Qu.: 1.7317
##  Median : 2.5915
##  Mean   : 2.4784
##  3rd Qu.: 3.0564
##  Max.   : 5.5829
```

We can see from the table that the median age of the patients was 65 years, and the mean of the log of their weights was 3.653. We may also summarise some aspects of this data graphically. The following plot shows the relationship between age of the patient and the log of the weight of the patient.

```r
plot(prostate$lcavol ~ prostate$age)
```



The above plot shows that their is mild evidence to suggest that the age of the patient is related to the log of the cancer volume.

Question 4

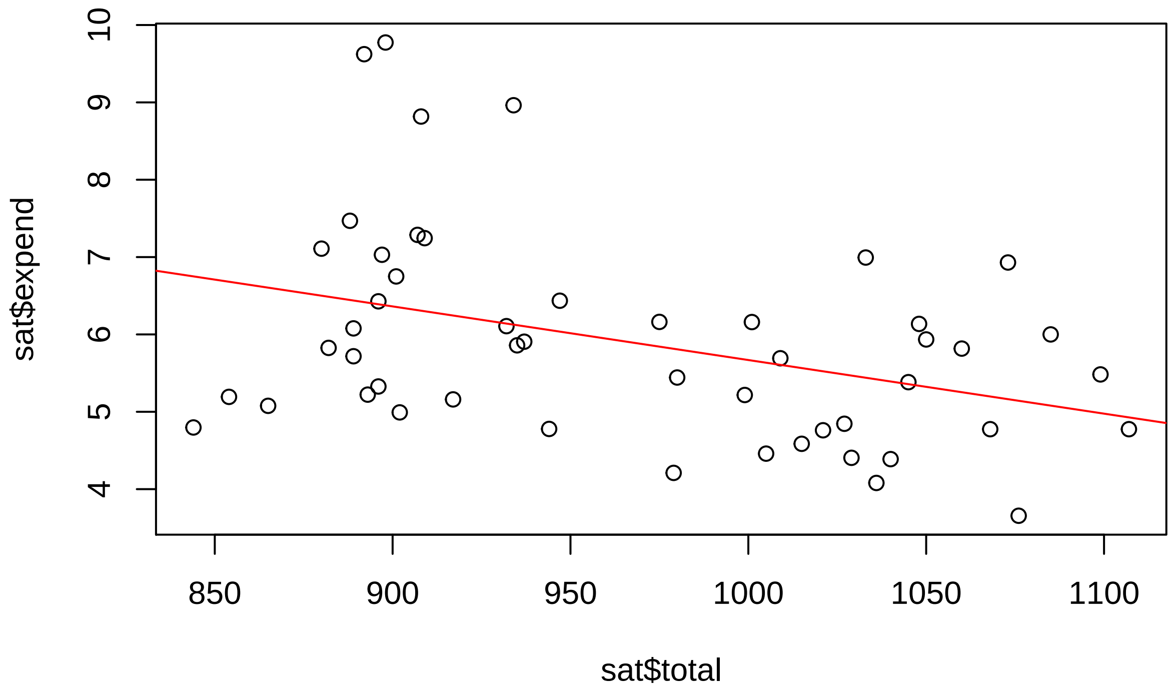
The following table represents a summary of data collected to study the expenditure of public schools.

```
data(sat)
summary(sat)
```

```
##      expend          ratio          salary          takers     
##  Min.   :3.656   Min.   :13.80   Min.   :25.99   Min.   : 4.00  
##  1st Qu.:4.882   1st Qu.:15.22   1st Qu.:30.98   1st Qu.: 9.00  
##  Median :5.768   Median :16.60   Median :33.29   Median :28.00  
##  Mean   :5.905   Mean   :16.86   Mean   :34.83   Mean   :35.24  
##  3rd Qu.:6.434   3rd Qu.:17.57   3rd Qu.:38.55   3rd Qu.:63.00  
##  Max.   :9.774   Max.   :24.30   Max.   :50.05   Max.   :81.00  
##      verbal           math           total       
##  Min.   :401.0   Min.   :443.0   Min.   : 844.0  
##  1st Qu.:427.2   1st Qu.:474.8   1st Qu.: 897.2  
##  Median :448.0   Median :497.5   Median : 945.5  
##  Mean   :457.1   Mean   :508.8   Mean   : 965.9  
##  3rd Qu.:490.2   3rd Qu.:539.5   3rd Qu.:1032.0  
##  Max.   :516.0   Max.   :592.0   Max.   :1107.0
```

We can see from this summary that the mean SAT score of the schools is 965.9, and that the mean expenditure per student of schools is $5905. We may also summarise some aspects of this data graphically. The following plot shows the relationship between expenditure per student and SAT test scores/

```
plot(sat$expend ~ sat$total)
abline(lm(sat$expend ~ sat$total), col="red")
```



This plot shows that there is actually a negative correlation between expenditure and SAT test scores which is surprising.

Question 5

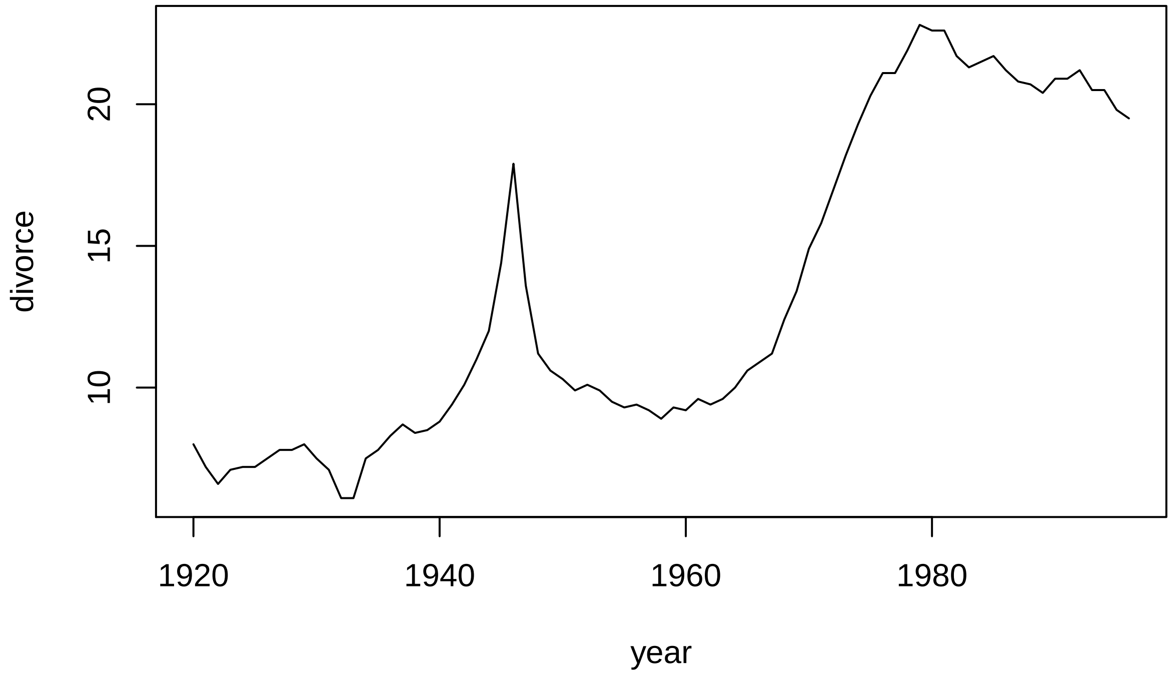
The following table represents a summary of data collected to study divorces in the US from 1920 to 1996.

```
data("divusa")
summary(divusa)
```

```
##       year         divorce        unemployed         femlab     
##  Min.   :1920   Min.   : 6.10   Min.   : 1.200   Min.   :22.70  
##  1st Qu.:1939   1st Qu.: 8.70   1st Qu.: 4.200   1st Qu.:27.47  
##  Median :1958   Median :10.60   Median : 5.600   Median :37.10  
##  Mean   :1958   Mean   :13.27   Mean   : 7.173   Mean   :38.58  
##  3rd Qu.:1977   3rd Qu.:20.30   3rd Qu.: 7.500   3rd Qu.:47.80  
##  Max.   :1996   Max.   :22.80   Max.   :24.900   Max.   :59.30  
##     marriage          birth           military     
##  Min.   : 49.70   Min.   : 65.30   Min.   : 1.940  
##  1st Qu.: 61.90   1st Qu.: 68.90   1st Qu.: 3.469  
##  Median : 74.10   Median : 85.90   Median : 9.102  
##  Mean   : 72.97   Mean   : 88.89   Mean   :12.365  
##  3rd Qu.: 80.00   3rd Qu.:107.30   3rd Qu.:14.266  
##  Max.   :118.10   Max.   :122.90   Max.   :86.641  
```

We can see from this table that the mean number of divorces was 13.27, and the number of women working varied from 22.7% to 59.3% during the time span. e may also summarise some aspects of this data graphically. The following plot shows how the divorce rate changed over time.

```
data <- data.frame(
  year = divusa$year,
  divorce = divusa$divorce
)
plot(data, type="l")
```



As the above plot shows, the number of divorces has fluctuated massively. It rapidly increased from 0 in 1930 to 17 in 1945 and then quickly decreased to 10 in around 1960. Starting in 1960 the divorce rate steadily increased up until 1980 where is began to level off at around 20.