STAT380: Assessment 2

Ash Midgley

March 23, 2016

Question 1

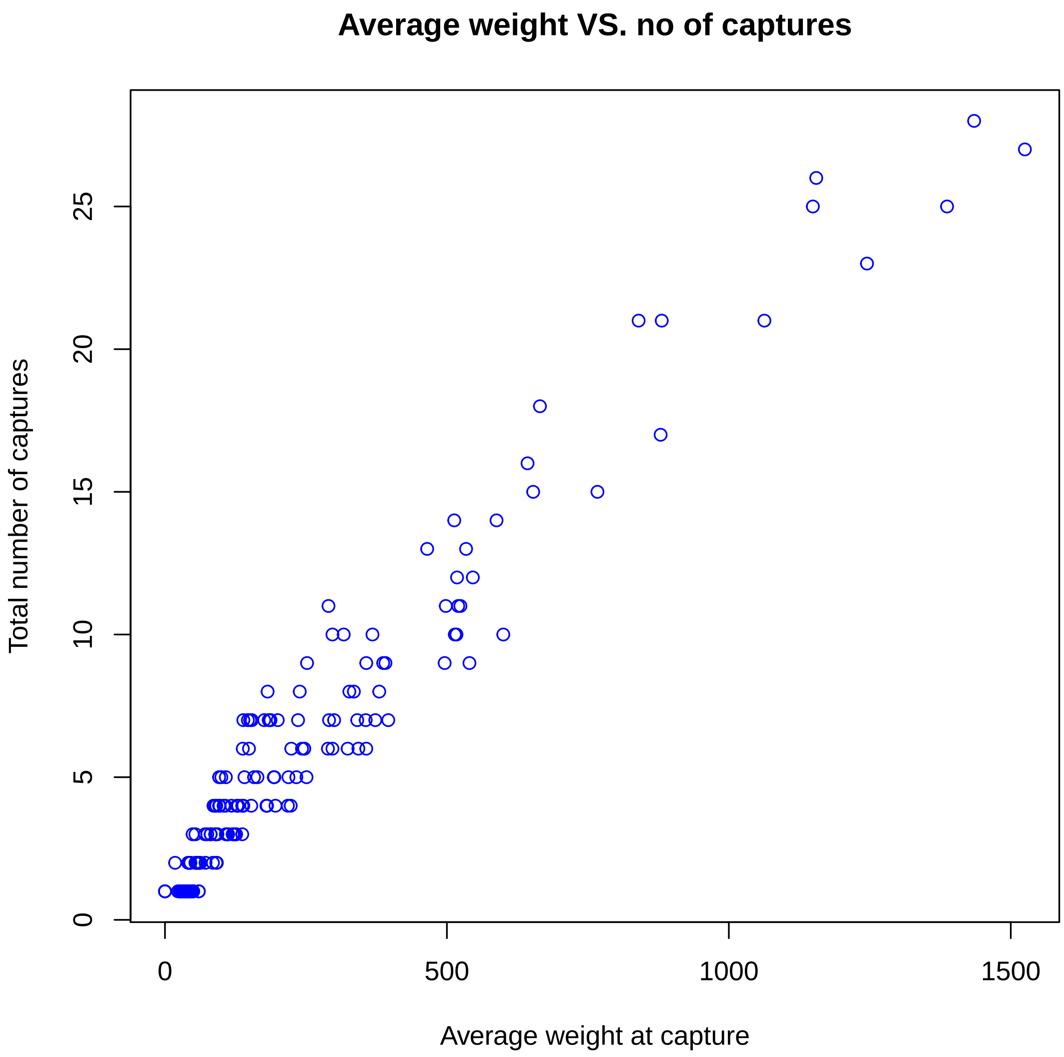
a)

The ’?’ symbol (question mark) is used to represent a missing value. The ’and’ symbol is used to delimit the data. The data does have a header. It is top line with the variables.

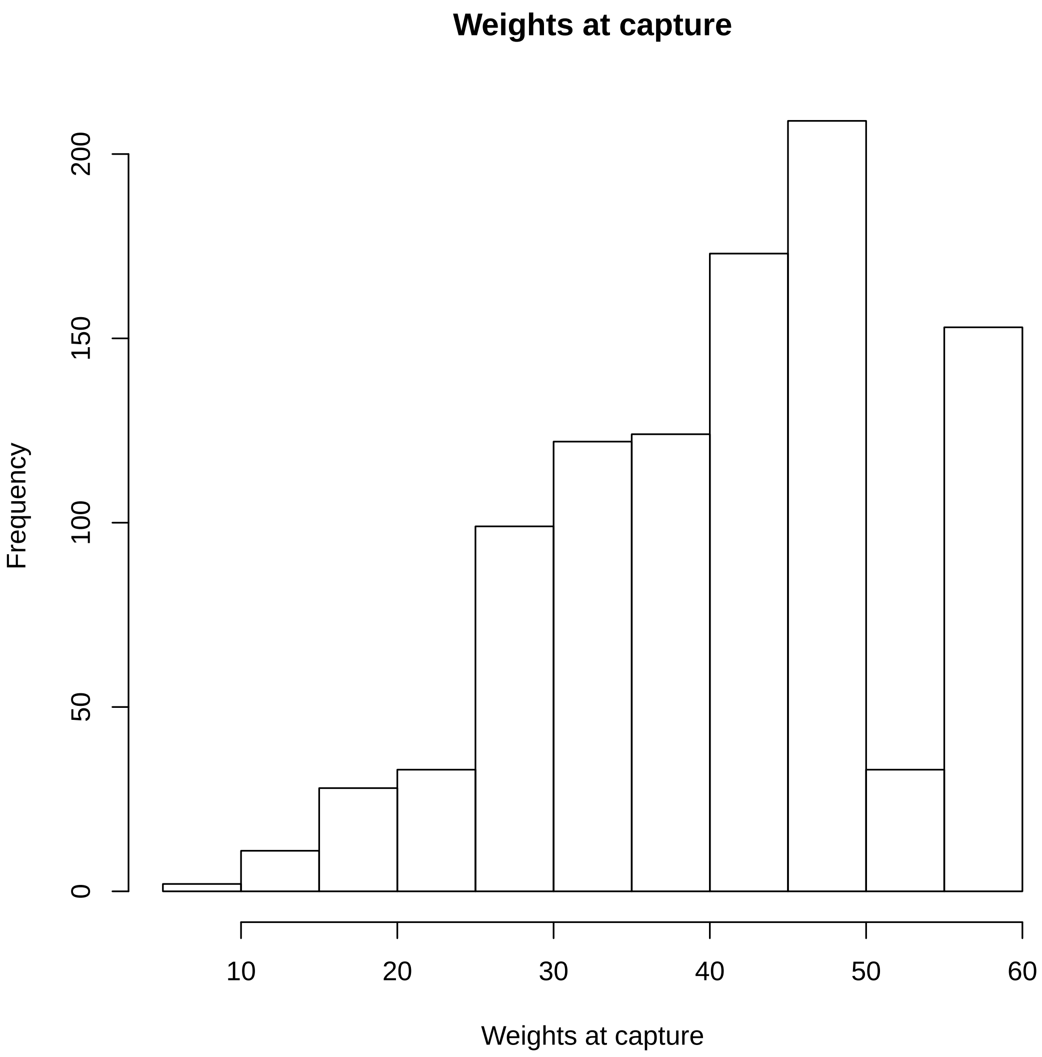
b)

```
my_data = read.table('vole380.txt', header=T, sep="&", na.strings = c("?"));
xvals = my_data[1:31]
zvals = my_data[32:60]
zvalsNum = apply(zvals, 1, as.numeric)
xones = c(1:173)
zmeans = c(1:173)
xrows = nrow(xvals)
xcols = ncol(xvals)
for(i in 1:xrows){
  onecount = 0
  for(k in 1:xcols){
    if(xvals[i, k] == 1){
      onecount = onecount+1
    }
  }
  xones[i] = onecount
}
```

```
zrows = nrow(zvals)

zcols = ncol(zvals)

for(i in 1:zrows){

  total=0

  numVals=0

  for(k in 1:zcols){

    if(!is.na(zvals[i,k])){

      total = total + zvals[i, k]

      numVals = numVals+1

    }

  }

  zmeans[i] = total

}

plot(zmeans, xones, main="Average weight VS. no of captures",

     ylab="Total number of captures",xlab="Average weight at capture",col="blue")
```



```
hist(zvalsNum ,main="Weights at capture",xlab="Weights at capture")
```



Generally the heavier the vole the more times it was captured (and vice versa).

c)

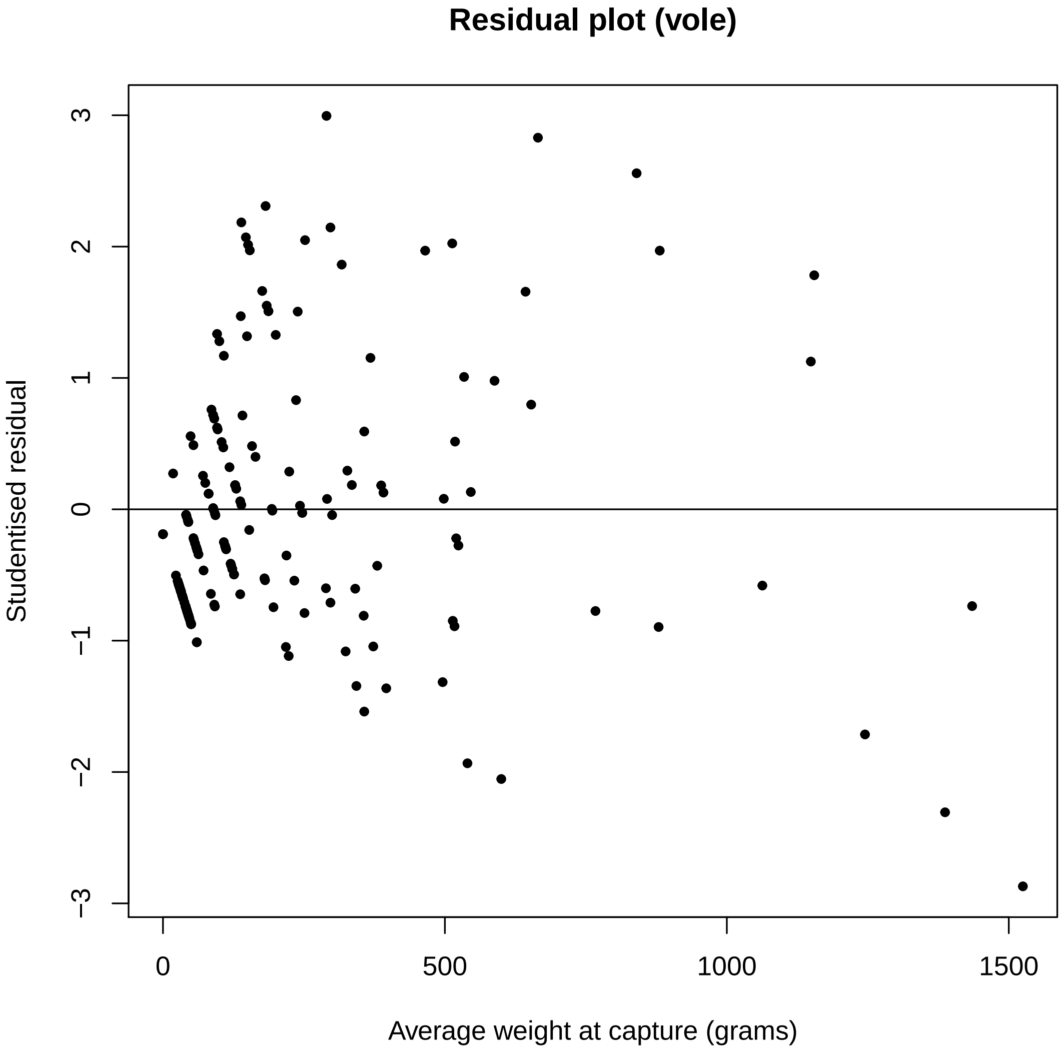
```
m1 = lm(xones~zmeans)
m1

##
## Call:
## lm(formula = xones ~ zmeans)
##
## Coefficients:
```

```
## (Intercept)       zmeans
##      1.26717      0.01932
```

```
resid1 = rstudent(m1)
confint(m1, level=0.95)
```

```
##                  2.5 %     97.5 %
## (Intercept) 0.98952819 1.54481394
## zmeans      0.01857017 0.02006257
```

```
plot(zmeans,resid1, pch=20, main="Residual plot (vole)",
     xlab="Average weight at capture (grams)", ylab="Studentised residual")
abline(h=0)
```



Yes, there evidence that that the regression assumptions have been violated as there is a trend in data on the left hand side of the plot.

Question 2

a)

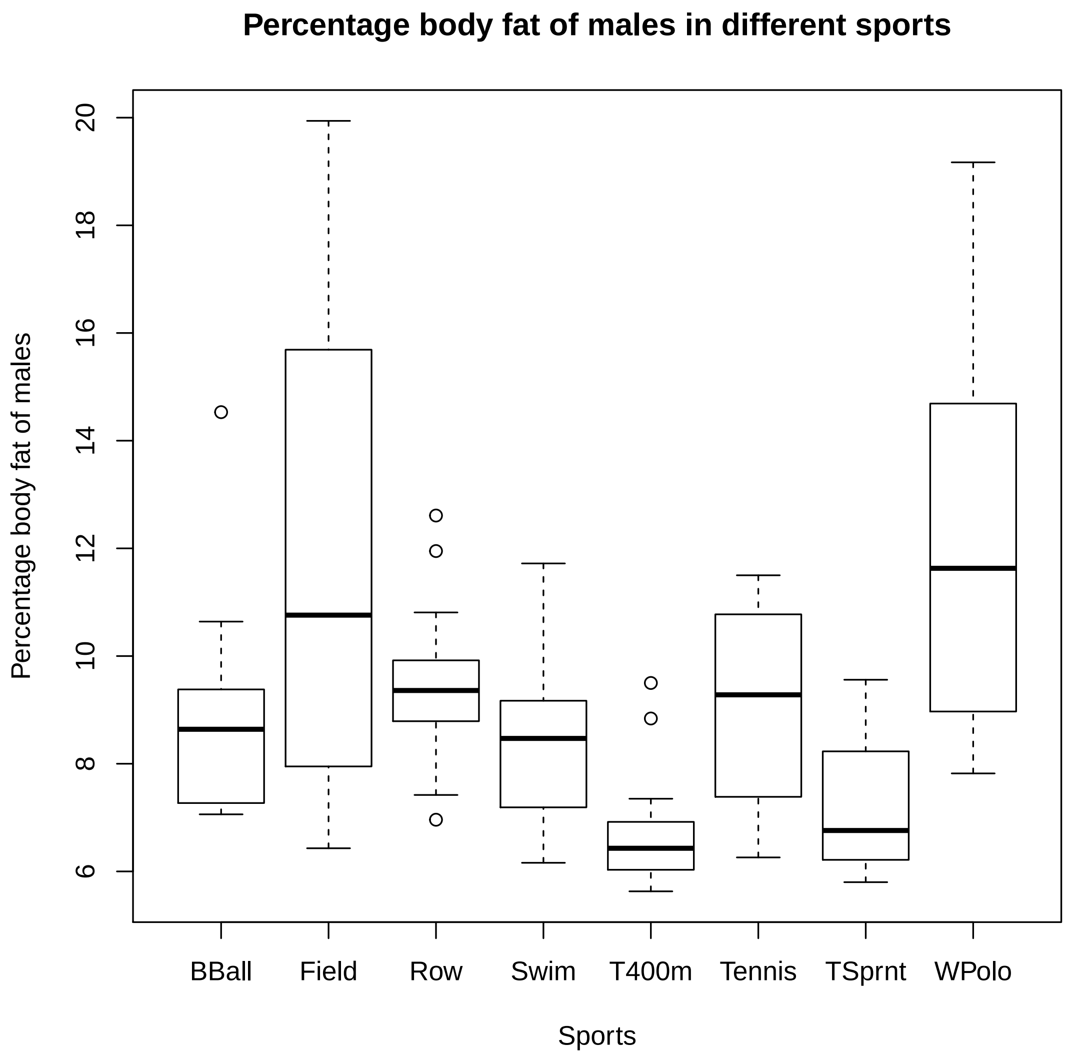
```
library(dplyr)

##
## 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

AIS_data = read.table('AIS.txt', header=T)
```

b)

```
d1 = AIS_data[AIS_data$Sex=="male",]
```

c)

```
d1$Sport <- as.character(d1$Sport)
d1$Sport <- as.factor(d1$Sport)
boxplot(d1$Bfat~d1$Sport, main="Percentage body fat of males in different sports",
        xlab="Sports", ylab="Percentage body fat of males")
```



d)

```
d2 = mutate(AIS_data, BMI=((AIS_data$Wt/(AIS_data$Ht/100)^2)))
```

e)

```
underweight = filter(d2, d2$BMI<18.5)

normal = filter(d2, d2$BMI>=18.5 & d2$BMI<25)

overweight = filter(d2, d2$BMI>=25 & d2$BMI<30)

obese = filter(d2, d2$BMI>30)
```

```
table(underweight$Sex)
```

```
##
## female   male
##      8      0
```

```
table(normal$Sex)
```

```
##
## female   male
##     79     73
```

```
table(overweight$Sex)
```

```
##
## female   male
##     12     24
```

```
table(obese$Sex)
```

```
##
## female   male
##      1      5
```

```
mean(underweight$Bfat)
```

```
## [1] 11.4425
```

```
mean(normal$Bfat)
```

```
## [1] 13.08388
```

```
mean(overweight$Bfat)
```

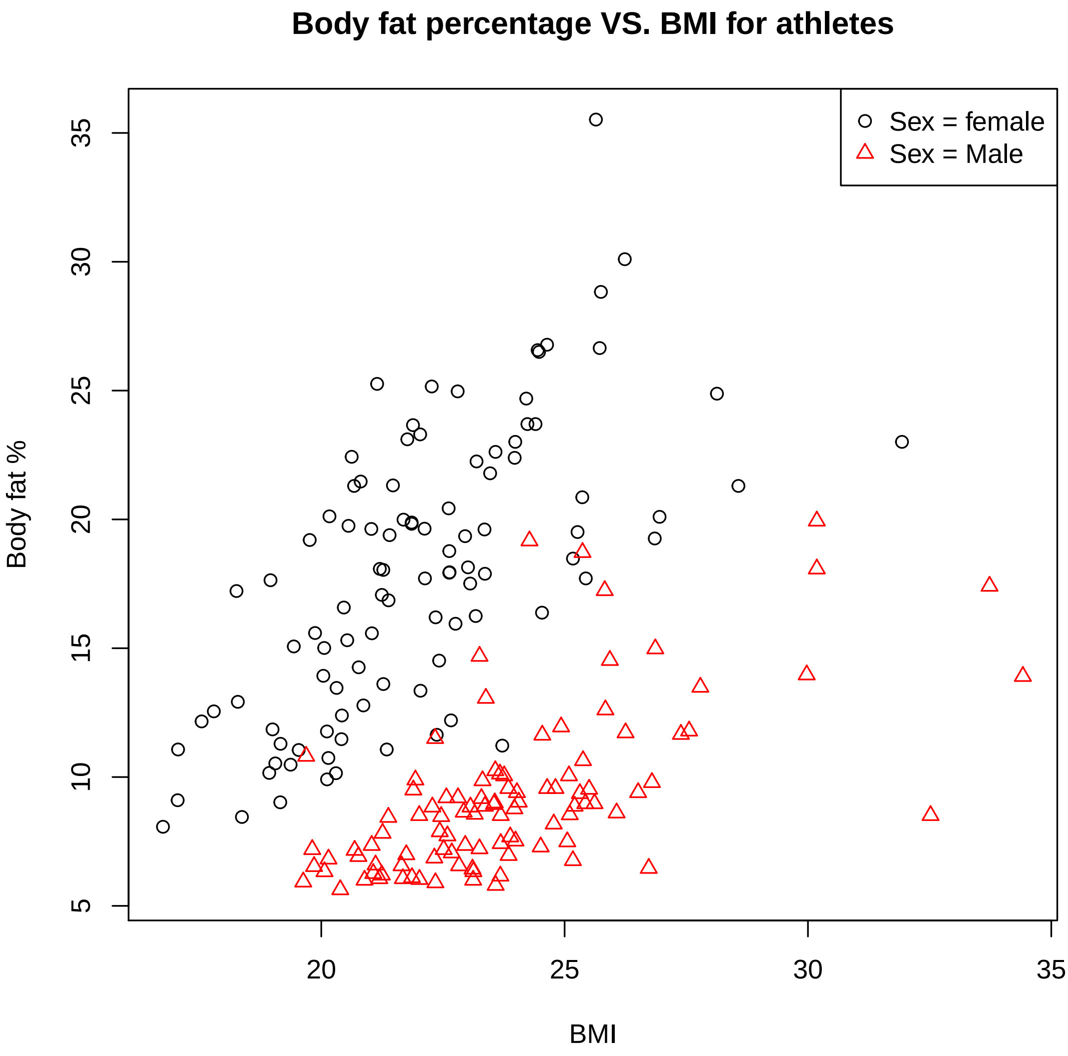
```
## [1] 15.20417
```

```
mean(obese$Bfat)
```

```
## [1] 16.81
```

f)

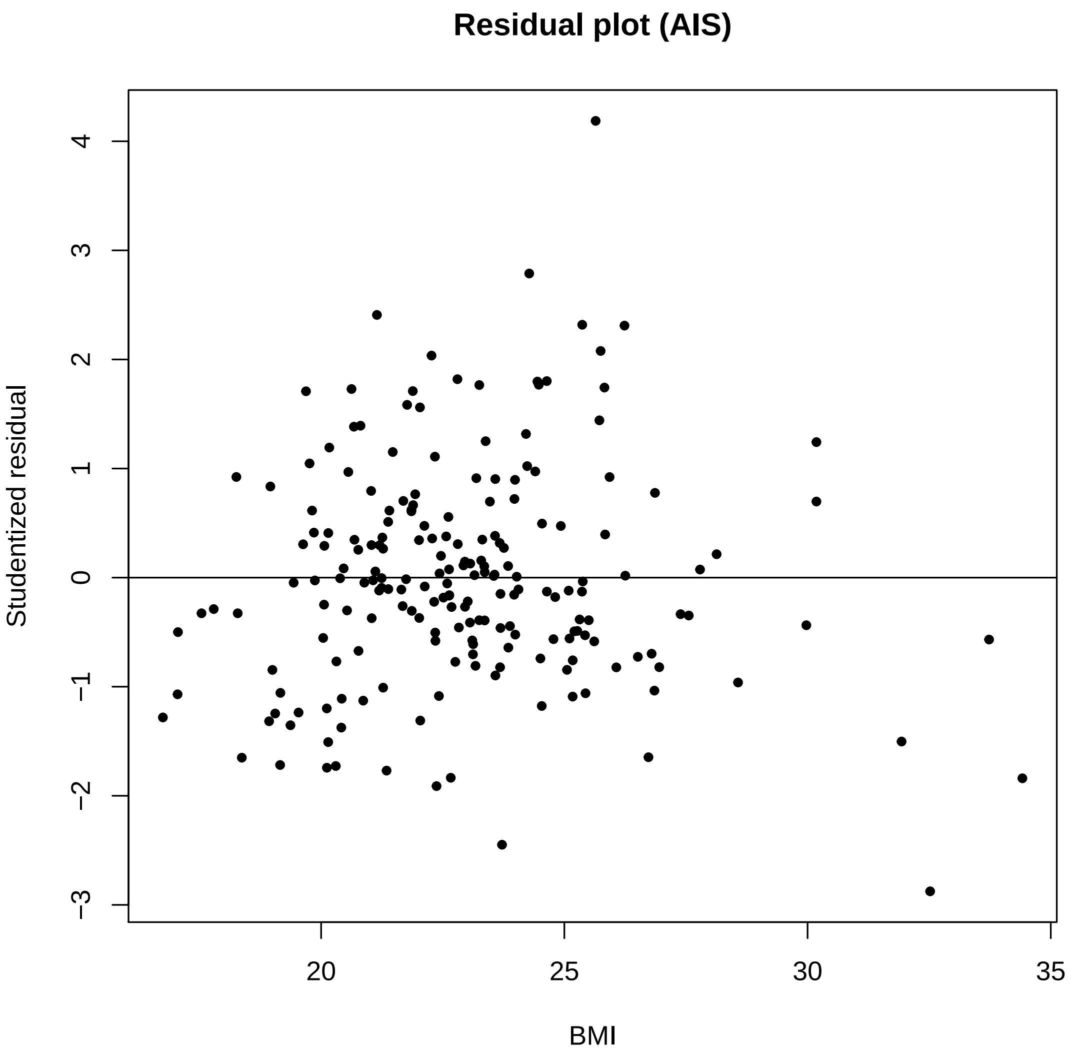
```r
plot(d2$BMI, d2$Bfat, pch=as.numeric(d2$Sex), col=as.numeric(d2$Sex),
     main="Body fat percentage VS. BMI for athletes", xlab="BMI", ylab="Body fat %")
legend("topright", legend=c('Sex = female', 'Sex = Male'), col= c(1,2), pch=c(1,2))
```



Males show to have a lower body fat percentage than females in the plot. Despite this, males appear to have a higher BMI score than females.

g)

```
m2 = lm(Bfat~BMI+Sex, data = d2)
s1 = summary(m2)
s1

##
## Call:
## lm(formula = Bfat ~ BMI + Sex, data = d2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.570 -2.015 -0.235  1.764 13.927
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -4.68185    2.03892  -2.296   0.0227 *
## BMI           1.02464    0.09135  11.216   <2e-16 ***
## Sexmale     -10.55988    0.52204 -20.228   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.495 on 199 degrees of freedom
## Multiple R-squared:  0.6843,Adjusted R-squared:  0.6811
## F-statistic: 215.7 on 2 and 199 DF,  p-value: < 2.2e-16

residm2 = rstudent(m2)
plot(d2$BMI,residm2,pch=20,main="Residual plot (AIS)",
     xlab="BMI",ylab="Studentized residual")
abline(h=0)
```



Yes there is slight evidence that the regression assumptions have been violated. Majority of the data fits the regression assumptions and shows no trends but Outliers skew the data to the left hand side of the plot.