Chapter 3 Exercises

#Exercise 1

Lets load in the data file and take a look:

```
dat=read.table("./ARM_Data/pyth/exercise2.1.dat",header=TRUE)
head(dat)
```

```
##       y   x1    x2
## 1 15.68 6.87 14.09
## 2  6.18 4.40  4.35
## 3 18.10 0.43 18.09
## 4  9.07 2.73  8.65
## 5 17.97 3.25 17.68
## 6 10.04 5.30  8.53
```

According to the text some of the data does not have y values; confirm that and create a subset of just the labelled data:

```
#how many rows have y label
print(paste("Labeled: ", sum(!is.na(dat$y))))
```

```
## [1] "Labeled:  40"
```

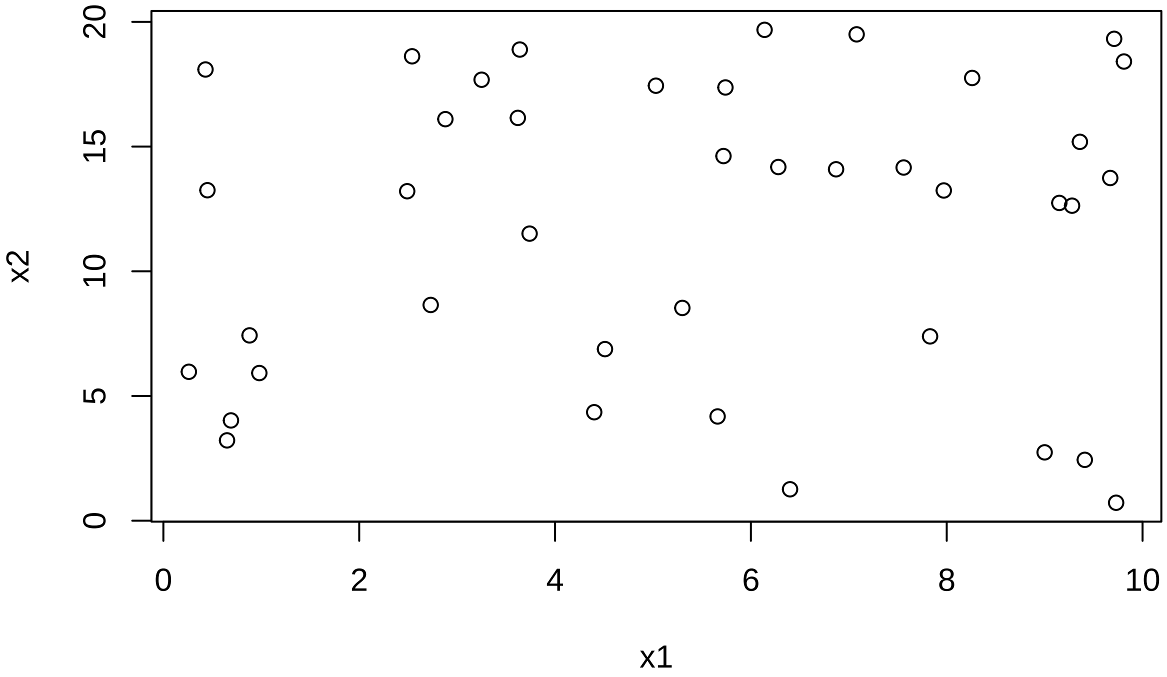
```
#how many rows do not have y label
print(paste("Unlabeled: ",sum(is.na(dat$y))))
```

```
## [1] "Unlabeled:  20"
```

```
label<-dat[!is.na(dat$y),]
```

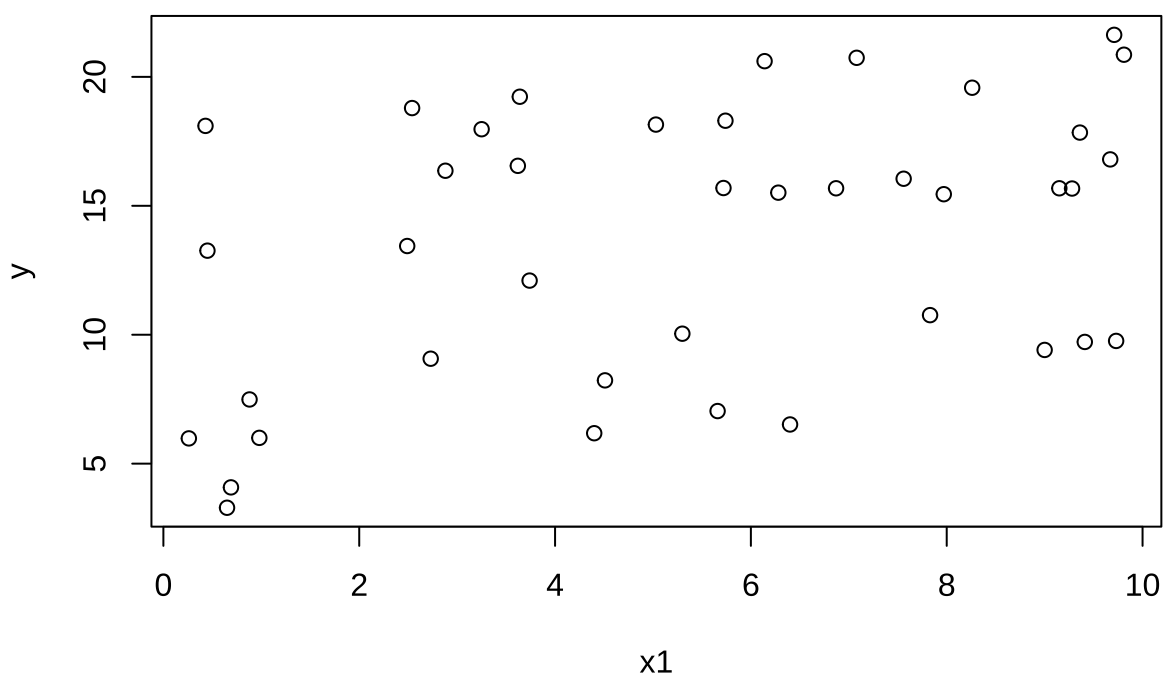
Lets take a look at the distribution of the predictors for the labelled data:

```
plot(label[c('x1','x2')])
```



Now lets plot the response against each predictor, starting with x1:

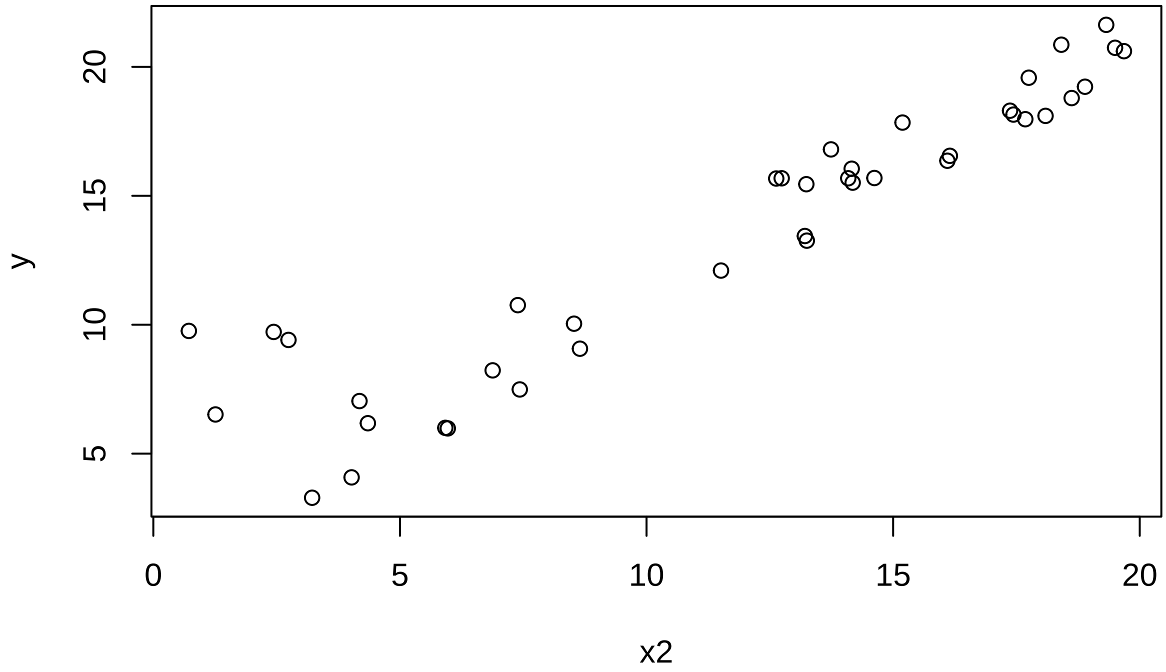
```
plot(label[c('x1','y')])
```



Not much of a relation on the margin…

Now x2:

```
plot(label[c('x2','y')])
```



Oh that looks quite a bit stronger.

Now lets fit a linear model to the labelled data:

```
fit = lm(label$y ~ label$x1 + label$x2)
summary(fit)
```

```
##
## Call:
## lm(formula = label$y ~ label$x1 + label$x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9585 -0.5865 -0.3356  0.3973  2.8548
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.31513    0.38769   3.392  0.00166 **
## label$x1     0.51481    0.04590  11.216 1.84e-13 ***
## label$x2     0.80692    0.02434  33.148  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9 on 37 degrees of freedom
## Multiple R-squared:  0.9724, Adjusted R-squared:  0.9709
## F-statistic: 652.4 on 2 and 37 DF,  p-value: < 2.2e-16
```