Short Course on R

Inferential statistics

The general linear model and its special cases

I. Correlation

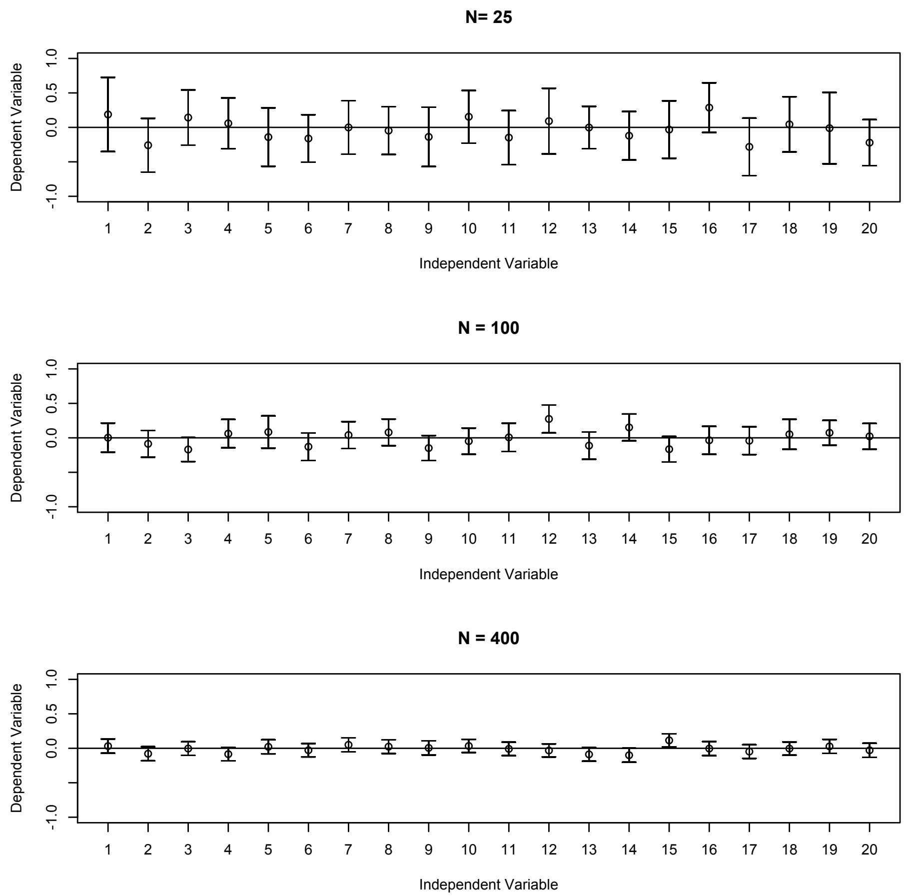
II. Multiple Regression

   A. Linear

   B. Logistic, Poisson, lognormal ...

III. t-tests and ANOVA

Confidence intervals, sample size, and Type I error



Confidence intervals

```
> op  <- par(mfrow=c(3,1))
> set.seed(42)
> x <- matrix(rnorm(500),ncol=20)
> error.bars(x,ylim=c(-1,1),main= "N= 25")
> abline(h=0)
>  x <- matrix(rnorm(2000),ncol=20)
>  error.bars(x,ylim=c(-1,1),main="N = 100")
>  abline(h=0)
>  x <- matrix(rnorm(8000),ncol=20)
>  error.bars(x,ylim=c(-1,1),main="N = 400")
> abline(h=0)
>  op <- par(mfrow=c(1,1))
```

Correlation

I. Testing a single correlation

II. Testing significance of many correlations

III. Testing the differences between correlations

   A. independent

   B. dependent

      1. same variables

      2. different variables

Finding correlations: cor

```
> data(sat.act)
> round(cor(sat.act,use="pairwise"),2)

          gender education   age   ACT  SATV  SATQ
gender      1.00      0.09 -0.02 -0.04 -0.02 -0.17
education   0.09      1.00  0.55  0.15  0.05  0.03
age        -0.02      0.55  1.00  0.11 -0.04 -0.03
ACT        -0.04      0.15  0.11  1.00  0.56  0.59
SATV       -0.02      0.05 -0.04  0.56  1.00  0.64
SATQ       -0.17      0.03 -0.03  0.59  0.64  1.00
```

Testing significance of a correlation: cor.test

```
> with(sat.act,cor.test(age,education))

   Pearson's product-moment correlation

data:  age and education
t = 17.3204, df = 698, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4942471 0.5980736
sample estimates:
      cor
0.5482695
```

```
> corr.test(sat.act)
Call:corr.test(x = sat.act)
Correlation matrix
          gender education   age   ACT  SATV  SATQ
gender      1.00      0.09 -0.02 -0.04 -0.02 -0.17
education   0.09      1.00  0.55  0.15  0.05  0.03
age        -0.02      0.55  1.00  0.11 -0.04 -0.03
ACT        -0.04      0.15  0.11  1.00  0.56  0.59
SATV       -0.02      0.05 -0.04  0.56  1.00  0.64
SATQ       -0.17      0.03 -0.03  0.59  0.64  1.00
Sample Size
          gender education age ACT SATV SATQ
gender       700       700 700 700  700  687
education    700       700 700 700  700  687
age          700       700 700 700  700  687
ACT          700       700 700 700  700  687
SATV         700       700 700 700  700  687
SATQ         687       687 687 687  687  687
Probability value
          gender education  age  ACT SATV SATQ
gender      0.00      0.02 0.58 0.33 0.62 0.00
education   0.02      0.00 0.00 0.00 0.22 0.36
age         0.58      0.00 0.00 0.00 0.26 0.37
ACT         0.33      0.00 0.00 0.00 0.00 0.00
SATV        0.62      0.22 0.26 0.00 0.00 0.00
SATQ        0.00      0.36 0.37 0.00 0.00 0.00
```

Testing many correlations

p values not corrected for multiple tests

Testing differences of correlations

```
> r.test(50,.3)    #test one correlation for significance
Correlation tests
Call:r.test(n = 50, r12 = 0.3)
Test of significance of a  correlation
 t value 2.18    with probability < 0.034
 and confidence interval  0.02 0.53
> r.test(30,.4,.6)       #test the difference between two independent
correlations
Correlation tests
Call:r.test(n = 30, r12 = 0.4, r34 = 0.6)
Test of difference between two independent correlations
 z value 0.99    with probability  0.32
> r.test(103,.4,.5,.1)   #Steiger case A (two dependent correlations
Correlation tests
Call:r.test(n = 103, r12 = 0.4, r34 = 0.5, r23 = 0.1)
Test of difference between two correlated  correlations
 t value -0.89    with probability < 0.37
> r.test(103,.5,.6,.7,.5,.5,.8)  #steiger Case B
Correlation tests
Call:r.test(n = 103, r12 = 0.5, r34 = 0.6, r23 = 0.7, r13 = 0.5, r14 =
0.5,
    r24 = 0.8)
Test of difference between two dependent correlations
 z value -1.2    with probability  0.23
```

Regression and multiple regression

I. The linear model (lm) for predicting one variable from another

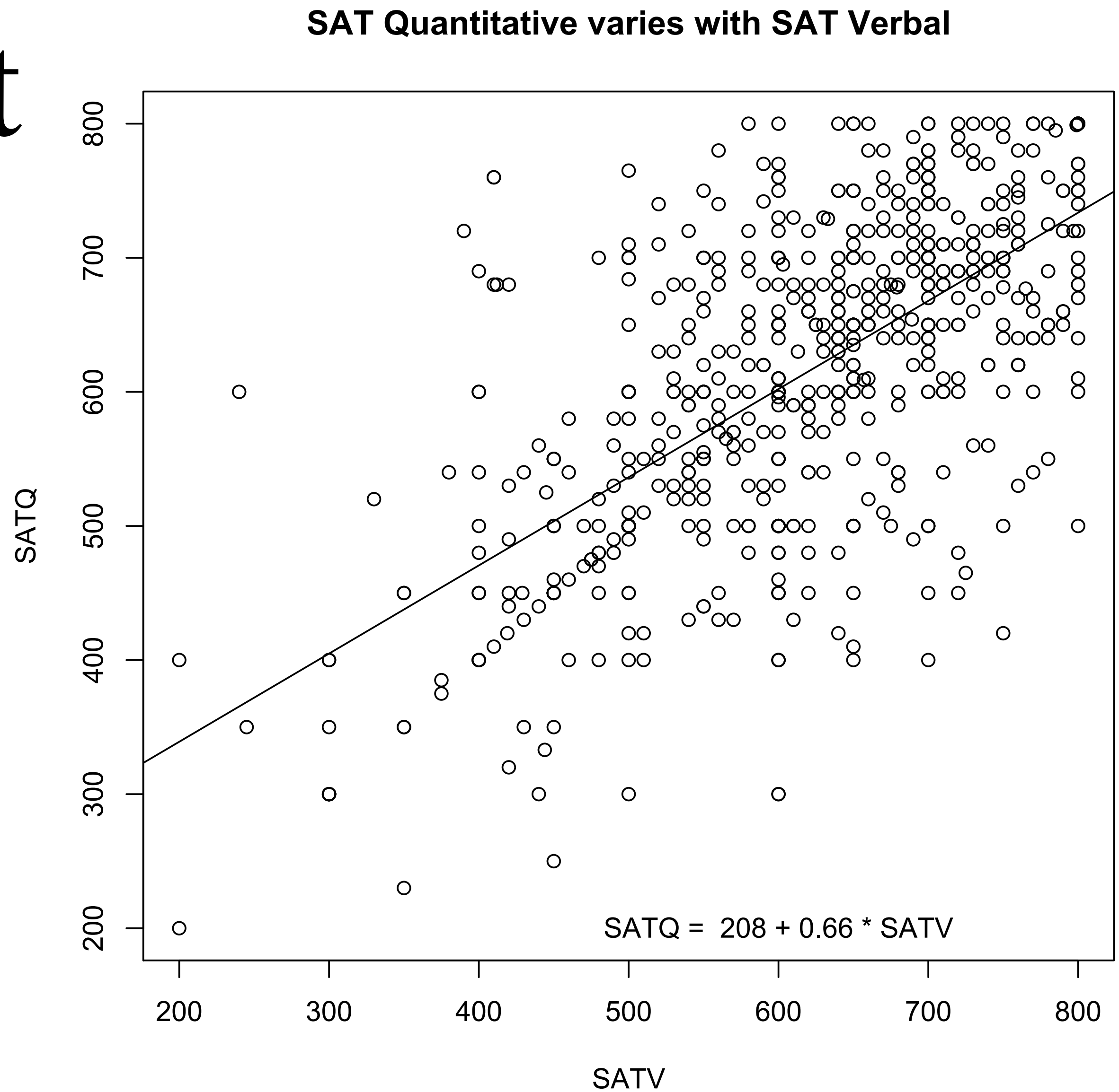
II. The linear model for predicting one variable from several

III. The linear model for predicting one variable from several including their interactions

Simple regression

```
> mod1 <- lm(SATQ ~ SATV,data=sat.act)
> summary(mod1)
Call:
lm(formula = SATQ ~ SATV, data = sat.act)
Residuals:
     Min       1Q   Median       3Q      Max 
-302.105  -46.477    2.403   51.319  282.845 
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 207.52528   18.57250   11.17   <2e-16 ***
SATV          0.65763    0.02983   22.05   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 88.5 on 685 degrees of freedom
  (13 observations deleted due to missingness)
Multiple R-squared: 0.4151,Adjusted R-squared: 0.4143 
F-statistic: 486.2 on 1 and 685 DF,  p-value: < 2.2e-16
```

And plot it




```
> with(sat.act,plot(SATQ~SATV,main="SAT Quantitative varies with SAT Verbal"))
> model = lm(SATQ~SATV,data=sat.act)
> abline(model)
> lab <- paste("SATQ = ",round(model$coef[1]),"+",round(model$coef[2],2),"* SATV")
> text(600,200,lab)
```

Multiple regression

```
> mod2 <- lm(SATQ ~ SATV  + gender,data=sat.act)
> summary(mod2)
Call:
lm(formula = SATQ ~ SATV + gender, data = sat.act)
Residuals:
     Min        1Q   Median        3Q       Max
-291.274  -50.457    5.635   51.891  295.343
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 269.89975   21.65705  12.462  < 2e-16 ***
SATV          0.65454    0.02925  22.375  < 2e-16 ***
gender      -36.80114    6.91400  -5.323 1.39e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
Residual standard error: 86.79 on 684 degrees of freedom
  (13 observations deleted due to missingness)
Multiple R-squared: 0.4384,Adjusted R-squared: 0.4367
F-statistic:   267 on 2 and 684 DF,  p-value: < 2.2e-16
```

Adding an interaction term

I. An interaction is asking does the effect of X on Y depend upon Z.

II. Can be found by correlating X*Z with Y

III. But, this product will be confounded with X and Z.

IV. Solution is to zero center X and Z.

Zero centering: the scale function

I. z <- scale(x) will convert to standard scores

II. w <- scale(x,scale=FALSE) just zero centers

III. scale returns a matrix, lm needs a data.frame

zero centering

original

```
>  headtail(sat.act,2,2)
      gender education age ACT SATV SATQ
29442      2         3  19  24  500  500
29457      2         3  23  35  600  500
...      ...       ... ... ...  ...  ...
39961      1         4  35  32  700  780
39985      1         5  25  25  600  600
> cent.data <- data.frame(scale(sat.act,scale=FALSE))
> z.data <- data.frame(scale(sat.act))
> headtail(z.data,2,2)
      gender education   age   ACT  SATV  SATQ
29442   0.74     -0.12 -0.69 -0.94 -0.99 -0.95
29457   0.74     -0.12 -0.27  1.34 -0.11 -0.95
...      ...       ...   ...   ...   ...   ...
39961  -1.35      0.59  0.99  0.72  0.78  1.47
39985  -1.35      1.29 -0.06 -0.74 -0.11 -0.09
```

z scored

```
> headtail(cent.data,2,2)
      gender education   age   ACT    SATV    SATQ
29442   0.35     -0.16 -6.59 -4.55 -112.23 -110.22
29457   0.35     -0.16 -2.59  6.45  -12.23 -110.22
...      ...       ...   ...   ...     ...     ...
39961  -0.65      0.84  9.41  3.45   87.77  169.78
39985  -0.65      1.84 -0.59 -3.55  -12.23  -10.22
```

centered

Interactions

```
> mod4 <- lm(SATQ ~ SATV  * gender,data=cent.data)
> summary(mod4)

Call:
lm(formula = SATQ ~ SATV * gender, data = cent.data)

Residuals:
     Min       1Q   Median       3Q      Max
-294.423  -49.876    5.577   53.210  291.100

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.26696    3.31211  -0.081    0.936
SATV          0.65398    0.02926  22.350  < 2e-16 ***
gender      -36.71820    6.91495  -5.310 1.48e-07 ***
SATV:gender  -0.05835    0.06086  -0.959    0.338
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 86.79 on 683 degrees of freedom
  (13 observations deleted due to missingness)
Multiple R-squared: 0.4391,  Adjusted R-squared: 0.4367
F-statistic: 178.3 on 3 and 683 DF,  p-value: < 2.2e-16
```

Interactions, incorrect main effects

```
> mod3 <- lm(SATQ ~ SATV  * gender,data=sat.act)
> summary(mod3)  #incorrect model

Call:
lm(formula = SATQ ~ SATV * gender, data = sat.act)

Residuals:
     Min       1Q   Median       3Q      Max 
-294.423  -49.876    5.577   53.210  291.100 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 211.19986   64.94501   3.252  0.00120 ** 
SATV          0.75009    0.10387   7.221 1.38e-12 ***
gender       -0.99528   37.98214  -0.026  0.97910    
SATV:gender  -0.05835    0.06086  -0.959  0.33804    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 86.79 on 683 degrees of freedom
  (13 observations deleted due to missingness)
Multiple R-squared: 0.4391,  Adjusted R-squared: 0.4367 
F-statistic: 178.3 on 3 and 683 DF,  p-value: < 2.2e-16
```

More detailed specifications

```
> mod5 <- lm(SATQ ~ SATV + ACT + gender*education,data=cent.data)
> summary(mod5)
Call:
lm(formula = SATQ ~ SATV + ACT + gender * education, data = cent.data
Residuals:
    Min      1Q  Median      3Q     Max 
-305.78  -46.07    5.67   51.82  261.21 
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        0.14552    3.10578   0.047    0.963    
SATV               0.46905    0.03306  14.187  < 2e-16 ***
ACT                7.86001    0.78567  10.004  < 2e-16 ***
gender           -34.07509    6.49943  -5.243 2.11e-07 ***
education         -2.56801    2.23493  -1.149    0.251    
gender:education  -5.45345    4.42642  -1.232    0.218    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 81.1 on 681 degrees of freedom
  (13 observations deleted due to missingness)
Multiple R-squared: 0.5117,  Adjusted R-squared: 0.5081 
F-statistic: 142.7 on 5 and 681 DF,  p-value: < 2.2e-16
```

Regressions from correlation matrix

I. Regression weights are function of covariance matrix, and can be calculated directly from that (or a correlation matrix)

II. Statistical tests can be applied if we know the sample size

III. Multiple analyses can be done at one time using the mat.regress function (psych)

mat.regress

```
> r <- cor(sat.act,use="pairwise")
> mat.regress(r,c(1:3),c(4:6))
$beta
            ACT  SATV  SATQ
gender    -0.05 -0.03 -0.18
education  0.14  0.10  0.10
age        0.03 -0.10 -0.09

$R
 ACT SATV SATQ
0.16 0.10 0.19

$R2
 ACT SATV SATQ
0.03 0.01 0.04
```

Comparisons of means

I. the t-test

   A. as a special case of the F-test

II. the F-test of Analysis of Variance

```
>  datafilename="http://personality-project.org/r/datasets/R.appendix1.data"
>  data.ex1=read.table(datafilename,header=T)   #read the data into a table
>  data.ex1
   Dosage Alertness
1       a        30
2       a        38
3       a        35
4       a        41
5       a        27
6       a        24
7       b        32
8       b        26
9       b        31
10      b        29
11      b        27
12      b        35
13      b        21
14      b        25
15      c        17
16      c        21
17      c        20
18      c        19
```

The data

for an ANOVA example

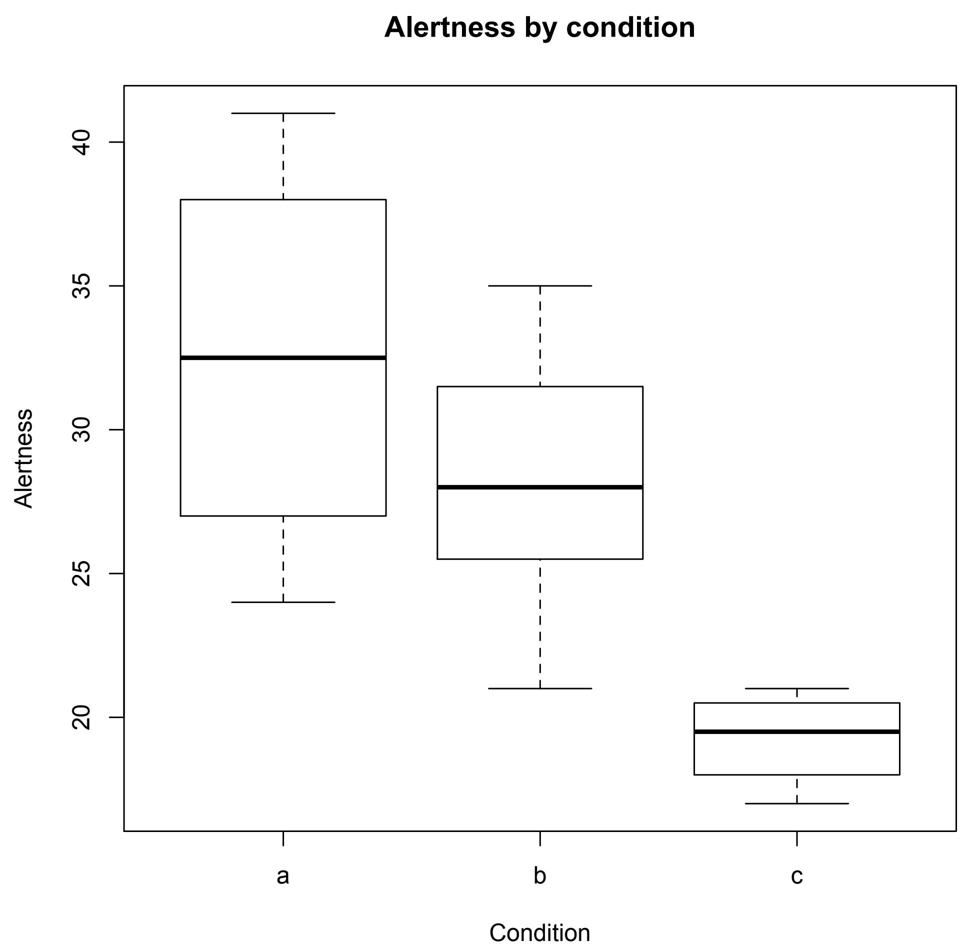
Select dose a and c

```
>  dose.2 <- subset(data.ex1,Dosage!="b")
>  t.test(Alertness~Dosage,data=dose.2)

   Welch Two Sample t-test

data:  Alertness by Dosage
t = 4.6907, df = 5.956, p-value = 0.003424
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
  6.325685 20.174315
sample estimates:
mean in group a mean in group c
          32.50           19.25
```

One way ANOVA

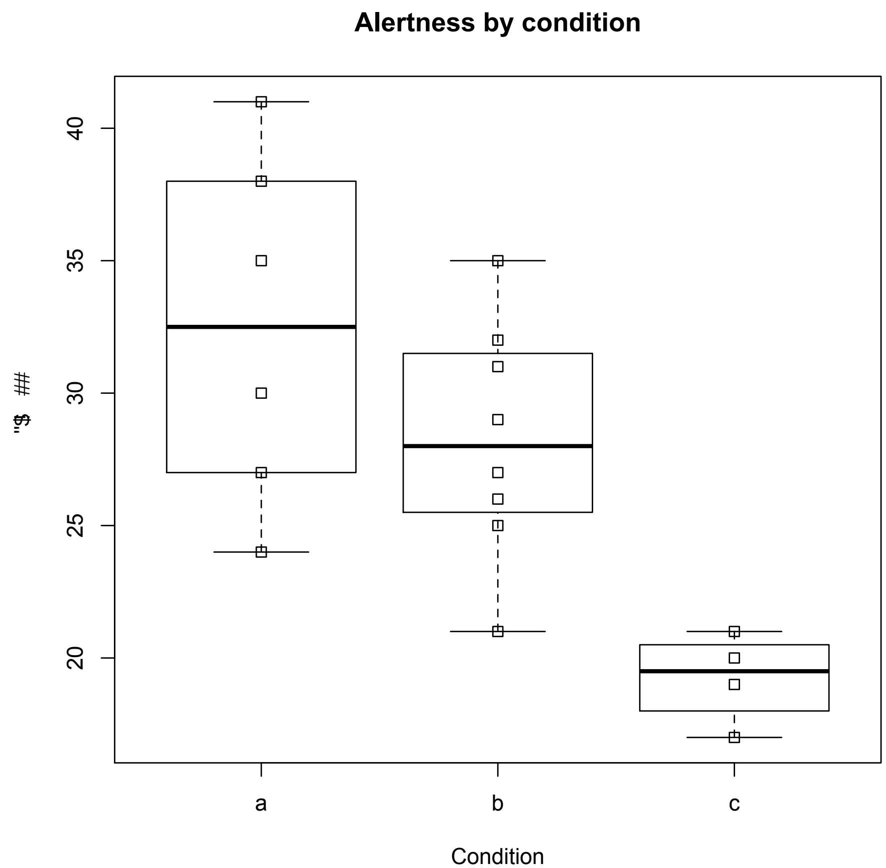
```
> aov.ex1 = aov(Alertness~Dosage,data=data.ex1)  #do the analysis of
variance
> summary(aov.ex1)              #show the summary table
            Df Sum Sq Mean Sq F value   Pr(>F)
Dosage       2 426.25  213.12  8.7887 0.002977 **
Residuals   15 363.75   24.25
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> print(model.tables(aov.ex1,"means"),digits=3)       #report the
means and the number of subjects/cell
Tables of means
Grand mean

27.66667

 Dosage
      a    b    c
   32.5 28.2 19.2
rep  6.0  8.0  4.0

> boxplot(Alertness~Dosage,data=data.ex1,main="Alertness by
condition",ylab="Alertness",xlab="Condition")        #graphical
summary appears in graphics window
```

Boxplot of results




```
> boxplot(Alertness~Dosage,data=data.ex1,main="Alertness by condition",ylab="Alertness",xlab="Condition")       #graphical summary appears in graphics window
```

Box + Stripchart

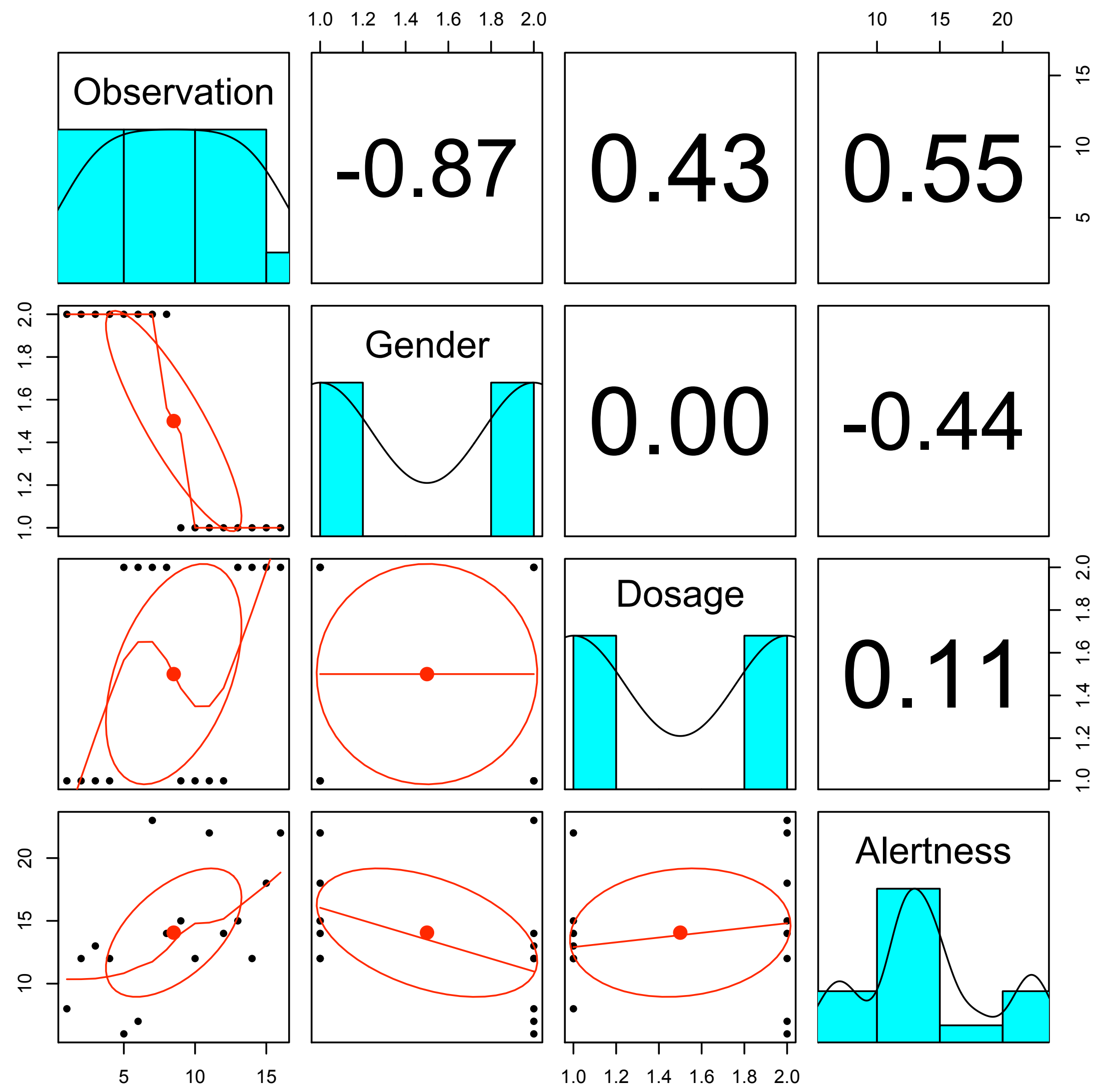



```
> boxplot(Alertness~Dosage,data=data.ex1,main="Alertness by
condition",ylab="Alertness",xlab="Condition")       #graphical summary appears in
graphics window
>
>  stripchart(Alertness~Dosage,data=data.ex1,vertical=TRUE,add=TRUE)
>
```

Two ANOVA

```
> datafilename="http://personality-project.org/R/datasets/R.appendix2.data"
> data.ex2=read.table(datafilename,header=T)   #read the data into a table
> data.ex2                                       #show the data
   Observation Gender Dosage Alertness
1            1      m      a         8
2            2      m      a        12
3            3      m      a        13
4            4      m      a        12
5            5      m      b         6
6            6      m      b         7
7            7      m      b        23
8            8      m      b        14
9            9      f      a        15
10          10      f      a        12
11          11      f      a        22
12          12      f      a        14
13          13      f      b        15
14          14      f      b        12
15          15      f      b        18
16          16      f      b        22
```

Possible confound?

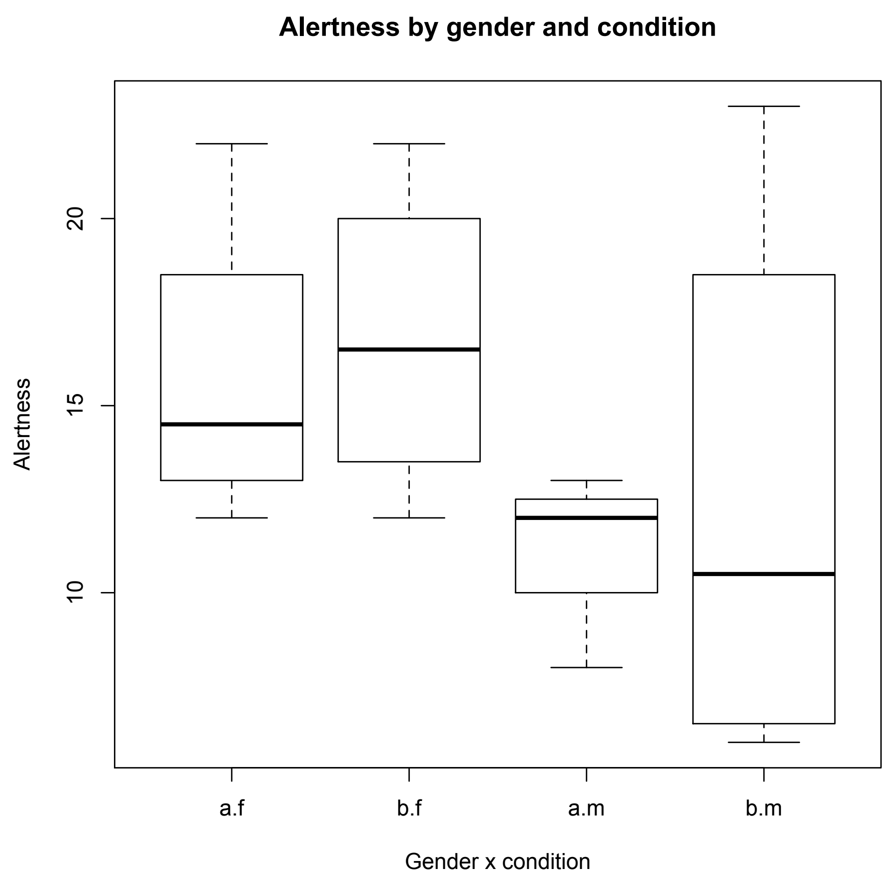


pairs.panels(data.ex2)

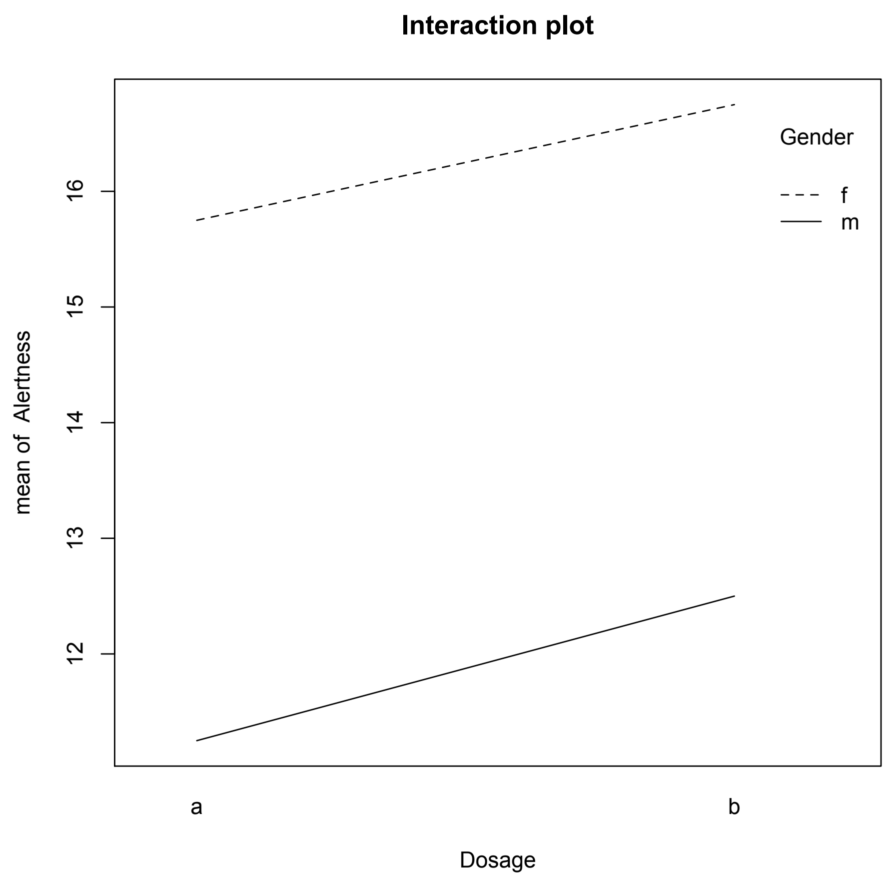
2 way ANOVA

```
> aov.ex2 = aov(Alertness~Gender*Dosage,data=data.ex2)       #do the
analysis of variance
> summary(aov.ex2)                                           #show the
summary table
              Df  Sum Sq Mean Sq F value Pr(>F)
Gender         1  76.562  76.562  2.9518 0.1115
Dosage         1   5.062   5.062  0.1952 0.6665
Gender:Dosage  1   0.063   0.063  0.0024 0.9617
Residuals     12 311.250  25.938
> print(model.tables(aov.ex2,"means"),digits=3)       #report the
means and the number of subjects/cell
Tables of means
Grand mean
14.0625
 Gender
Gender
    f     m
16.25 11.88
 Dosage
Dosage
    a     b
13.50 14.62
```

```
 Gender:Dosage
      Dosage
Gender a     b
     f 15.75 16.75
     m 11.25 12.50
```


Alertness by gender and condition
Alertness
10
15
20
a.f
b.f
a.m
b.m
Gender x condition

An interaction plot




```
 with(data.ex2,
interaction.plot(Dosage,Gender,Alertness,main="Interaction plot"))
```

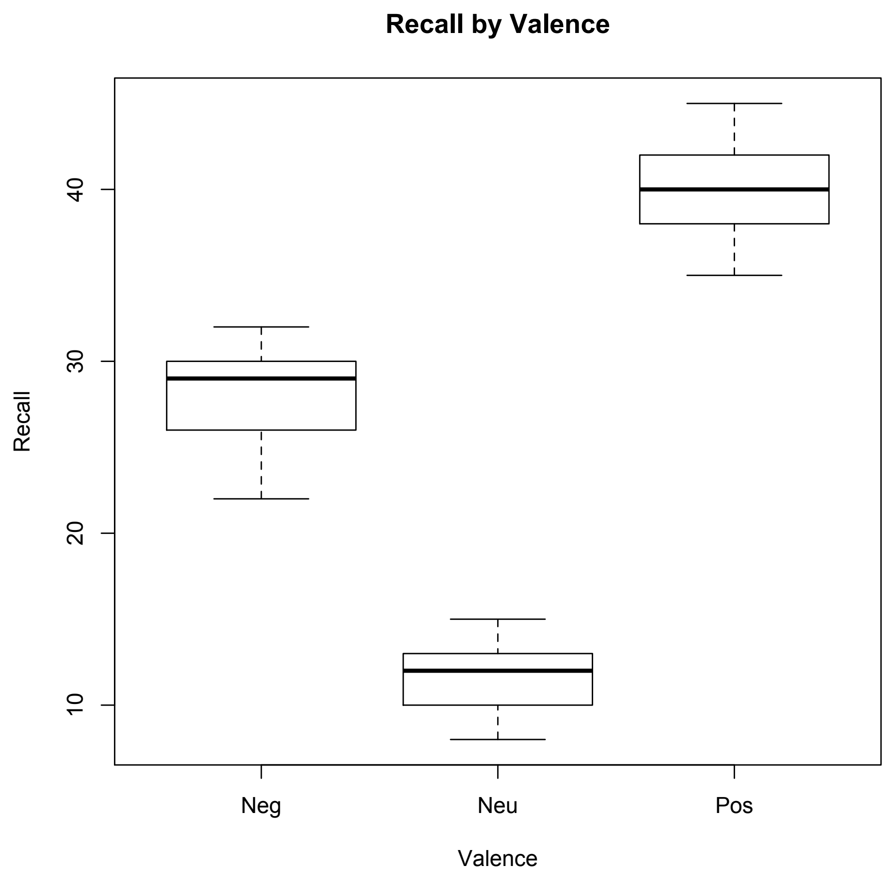
One way, repeated measures

```
> datafilename="http://personality-project.org/r/datasets/
R.appendix3.data"
> data.ex3=read.table(datafilename,header=T)   #read the data into a
table
> data.ex3                                          #show the data
   Observation Subject Valence Recall
1            1     Jim     Neg     32
2            2     Jim     Neu     15
3            3     Jim     Pos     45
4            4  Victor     Neg     30
5            5  Victor     Neu     13
6            6  Victor     Pos     40
7            7    Faye     Neg     26
8            8    Faye     Neu     12
9            9    Faye     Pos     42
10          10     Ron     Neg     22
11          11     Ron     Neu     10
12          12     Ron     Pos     38
13          13   Jason     Neg     29
14          14   Jason     Neu      8
15          15   Jason     Pos     35
```

Repeated measures ANOVA

```
> aov.ex3 = aov(Recall~Valence+Error(Subject/Valence),data.ex3)
> summary(aov.ex3)
Error: Subject
          Df  Sum Sq Mean Sq F value Pr(>F)
Residuals  4 105.067  26.267
Error: Subject:Valence
          Df  Sum Sq Mean Sq F value    Pr(>F)
Valence    2 2029.73 1014.87  189.11 1.841e-07 ***
Residuals  8   42.93    5.37
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> print(model.tables(aov.ex3,"means"),digits=3)       #report the
means and the number of subjects/cell
Tables of means
Grand mean
26.46667
 Valence
Valence
 Neg  Neu  Pos
27.8 11.6 40.0
```

Plotting the results

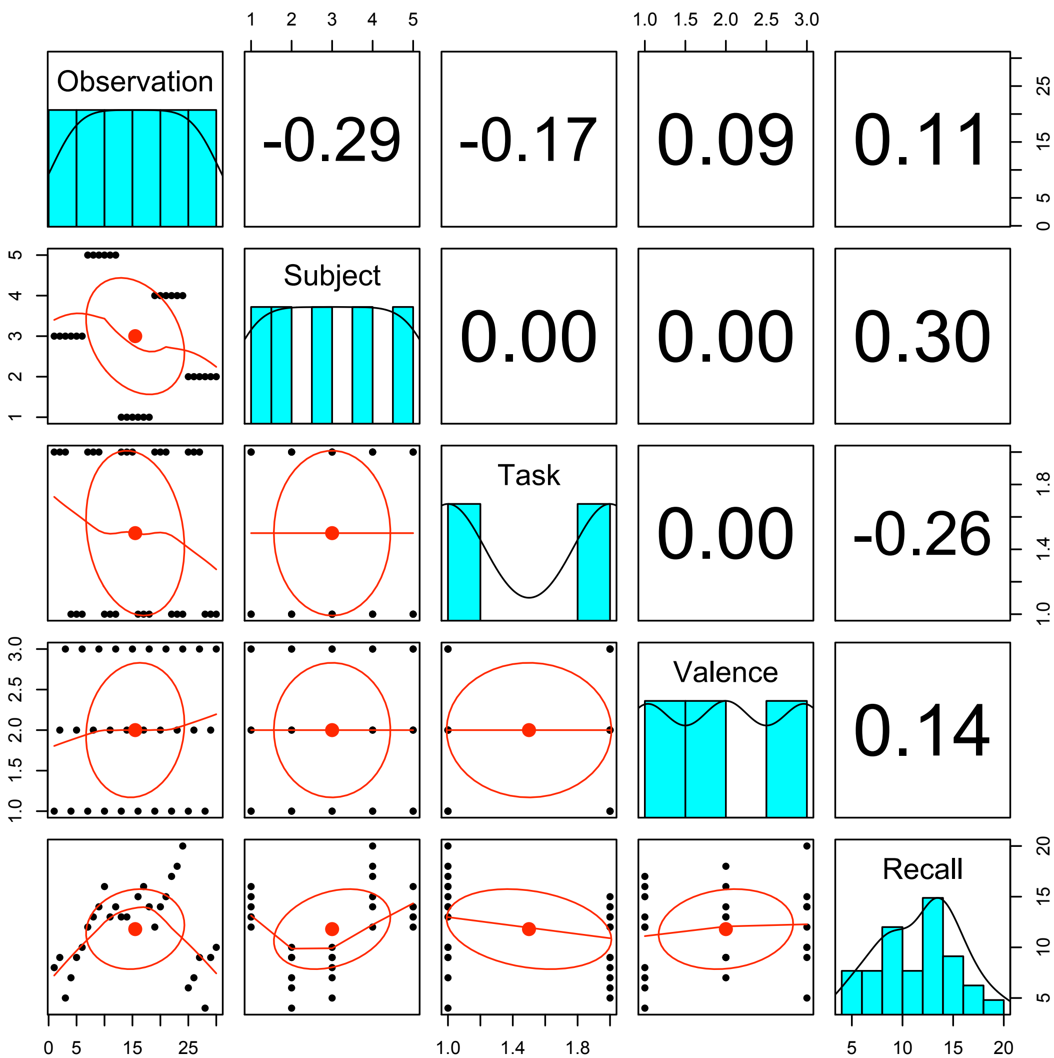



```
>  boxplot(Recall~Valence,data=data.ex3,main="Recall by
Valence",xlab="Valence",ylab="Recall")          #graphical output
```

Two way repeated ANOVA

```
> datafilename="http://personality-project.org/r/datasets/R.appendix4.data"
> data.ex4=read.table(datafilename,header=T)   #read the data into a table
>  data.ex4                                        #show the data
   Observation Subject Task Valence Recall
1            1     Jim Free     Neg      8
2            2     Jim Free     Neu      9
3            3     Jim Free     Pos      5
4            4     Jim Cued     Neg      7
5            5     Jim Cued     Neu      9
6            6     Jim Cued     Pos     10
7            7  Victor Free     Neg     12
8            8  Victor Free     Neu     13
9            9  Victor Free     Pos     14
10          10  Victor Cued     Neg     16
11          11  Victor Cued     Neu     13
12          12  Victor Cued     Pos     14
13          13    Faye Free     Neg     13
14          14    Faye Free     Neu     13
15          15    Faye Free     Pos     12
16          16    Faye Cued     Neg     15
17          17    Faye Cued     Neu     16
18          18    Faye Cued     Pos     14
19          19     Ron Free     Neg     12
20          20     Ron Free     Neu     14
21          21     Ron Free     Pos     15
22          22     Ron Cued     Neg     17
23          23     Ron Cued     Neu     18
24          24     Ron Cued     Pos     20
25          25   Jason Free     Neg      6
26          26   Jason Free     Neu      7
27          27   Jason Free     Pos      9
28          28   Jason Cued     Neg      4
29          29   Jason Cued     Neu      9
30          30   Jason Cued     Pos     10
```

Design is clean

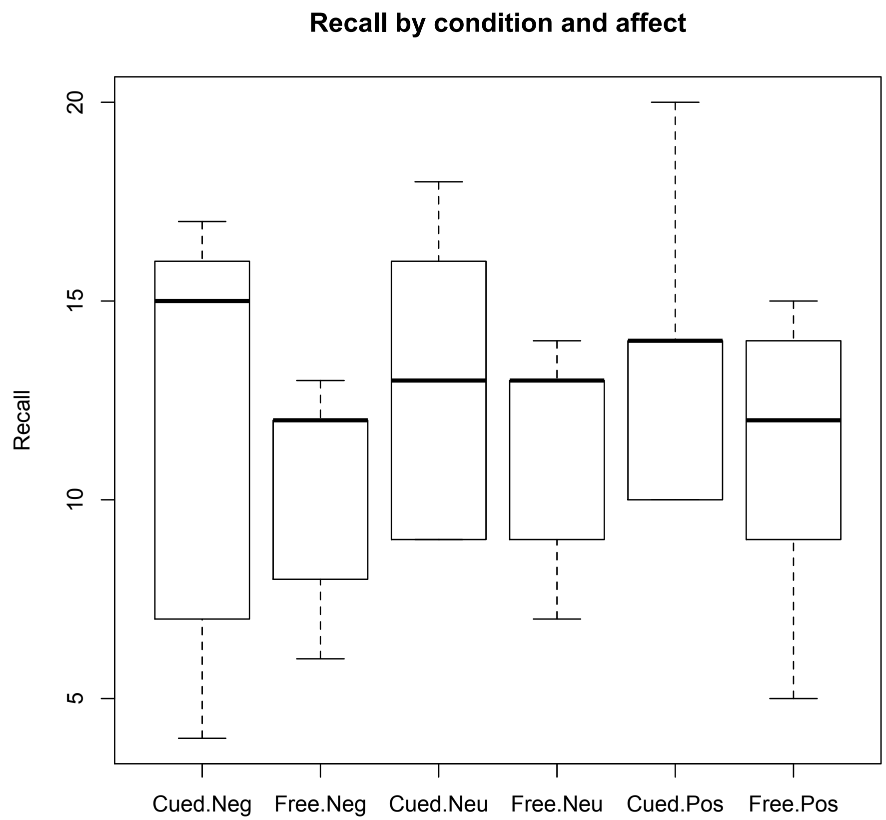


2 way repeated anova

```
>  aov.ex4=aov(Recall~(Task*Valence)+Error(Subject/
(Task*Valence)),data.ex4 )
>
> summary(aov.ex4)

Error: Subject
          Df Sum Sq Mean Sq F value Pr(>F)
Residuals  4 349.13   87.28

Error: Subject:Task
          Df  Sum Sq Mean Sq F value  Pr(>F)
Task       1 30.0000 30.0000  7.3469 0.05351 .
Residuals  4 16.3333  4.0833
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Error: Subject:Valence
          Df  Sum Sq Mean Sq F value Pr(>F)
Valence    2  9.8000  4.9000  1.4591 0.2883
Residuals  8 26.8667  3.3583

Error: Subject:Task:Valence
             Df  Sum Sq Mean Sq F value Pr(>F)
Task:Valence  2  1.4000  0.7000  0.2907 0.7553
Residuals     8 19.2667  2.4083
```

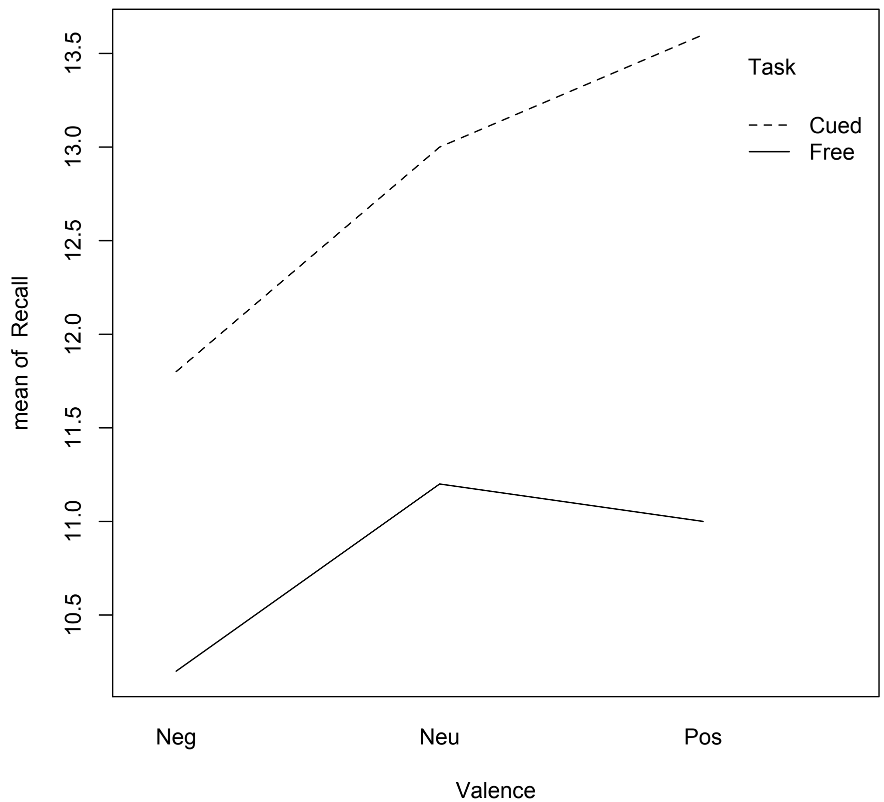
The means

```
> print(model.tables(aov.ex4,"means"),digits=3)       #report the 
means and the number of subjects/cell
Tables of means
Grand mean
11.8

 Task
Task
Cued Free 
12.8 10.8 

 Valence
Valence
 Neg  Neu  Pos 
11.0 12.1 12.3 

 Task:Valence
      Valence
Task   Neg  Neu  Pos 
  Cued 11.8 13.0 13.6
  Free 10.2 11.2 11.0
```




```
> boxplot(Recall~Task*Valence,data=data.ex4,main="Recall by condition
and affect",ylab="Recall") #graphical summary of means of the 6 cells
```

Interaction plots




```
with(data.ex4,interaction.plot(Valence,Task,Recall))    #another way
to graph the interaction
```