Psychology 205: Research Methods in Psychology

Using R to analyze the data for study 2

Department of Psychology
Northwestern University
Evanston, Illinois USA

November, 2012

Outline

1. Getting ready
2. Data analysis with R
   - Descriptive analysis
   - Linear modeling
   - Graphic displays of linear model effects
3. ANOVA
4. "Simple" effects
5. Advanced tricks
   - Recoding the data
   - Plotting ANOVA values

Steps for data analysis

- Install software on your computer or locate computer with software
  - (e.g., R, systat, SPSS)
- Prepare data for analysis
  - Subjects (rows) x variables (columns)
  - first row contains labels
- Read data
  - either copy and read.clipboard, or read a file

Using R

- Install R (download from http://cran.r-project.org)
  - Choose your appropriate operating system
  - Follow the installation directions
- run R
- add *psych* package using the package installer
  - on a PC, click on package options and install packages
  - on a Mac, go to package installer and get list, choose psych
  - or, just `install.packages("psych")`
- Remember R is an introverted and somewhat obsessive compulsive program.
  - It will not volunteer information unless you ask
  - It is sensitive to spelling errors
  - But it is very patient and very powerful. It will do anything you ask (if you ask politely).

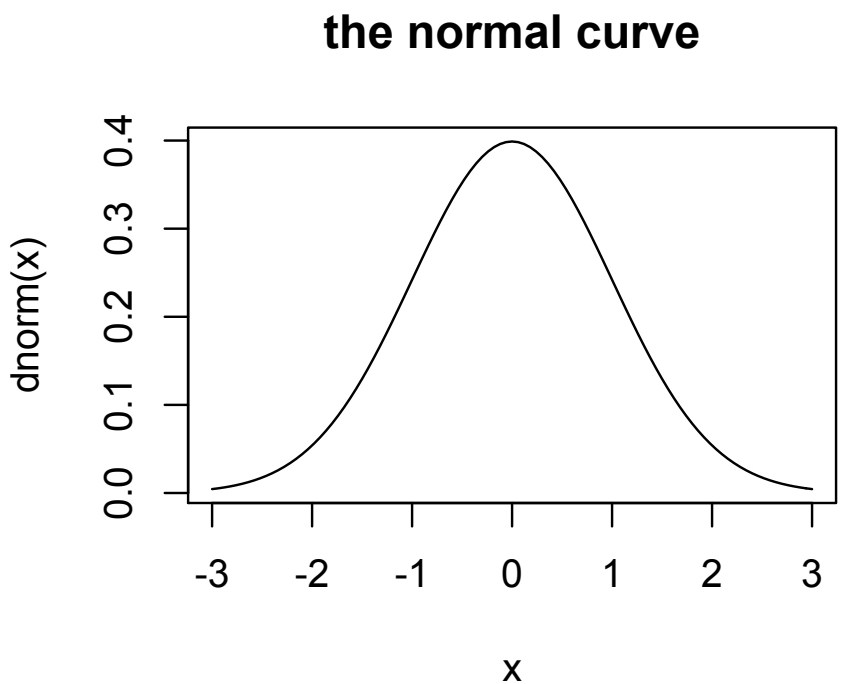
R is just a fancy desk calculator

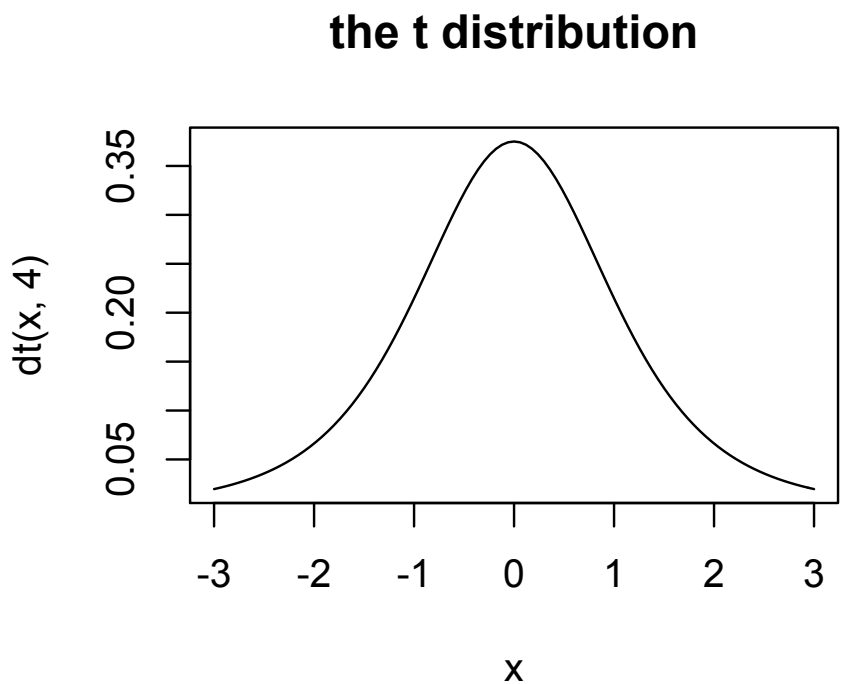
It will
Add
Multiply
Exponentiate
follow directions explicitly
treat data as vectors

```
> 2+3
[1] 5
> 3*4
[1] 12
> 2^5
[1] 32
> exp(1:4)
[1]  2.718282  7.389056 20.085537 54.598150
> round(exp(1:4))
[1]  3  7 20 55
> round(exp(1:4),2)
[1]  2.72  7.39 20.09 54.60
> x <- 2:6
> y <- 3:7
> x
[1] 2 3 4 5 6
> y
[1] 3 4 5 6 7
> x*y
[1]  6 12 20 30 42
```

R is also a very fancy graphical device

```
op <- par(mfrow=c(1,2))
curve(dnorm(x),-3,3,main="the normal curve")
curve(dt(x,4),-3,3,main="the t distribution")
```





R is a statistical look up table

```
> pnorm(1.96)
[1] 0.9750021
> pnorm(1.96) #Probability of a normal < 1.96
[1] 0.9750021
> pt(1.96,22) #probability of a t with df = 22 < 1.96
[1] 0.9686083
> pt(2,40) #probability of a t with df=40 < 2
[1] 0.9738388
> pf(4,1,40) #probability of F(1,40) < 4
[1] 0.9476777
> dbinom(0:10,10,.5) * 1024 #binomial distribution
 [1]   1  10  45 120 210 252 210 120  45  10   1
```

R will block randomize for you

```
library(psych)  #make sure that you have already installed it
#note that you need to say library(psych) the first time after starting \R{}
cond <- block.random(16,c(drug=2,time=2))  #specify two variables
cond  #show the resulting output
```

```
    blocks drug time
S1       1    2    2
S2       1    1    1
S3       1    1    2
S4       1    2    1
S5       2    2    1
S6       2    2    2
S7       2    1    1
S8       2    1    2
S9       3    1    2
S10      3    2    2
S11      3    1    1
S12      3    2    1
S13      4    1    2
S14      4    1    1
S15      4    2    1
S16      4    2    2
```

Consider the simulation experiment

- http://personality-project.org/revelle/syllabi/205/simulation/simulation.experiment.php
  - Not a good way to do an experiment, but will produce data for examples
  - Chose to generate 100 completely "random" subjects
- http://personality-project.org/revelle/syllabi/205/simulation/simulation.specification.php
- my results at http://personality-project.org/revelle/syllabi/205/simulation/simulating.personality.results.php
  - These results will differ each time I do the experiment
  - Results will differ as a function of conditions

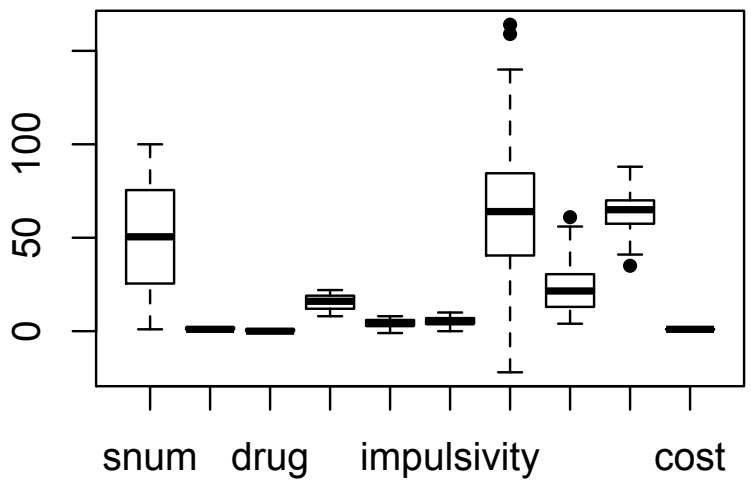
Data analysis with R

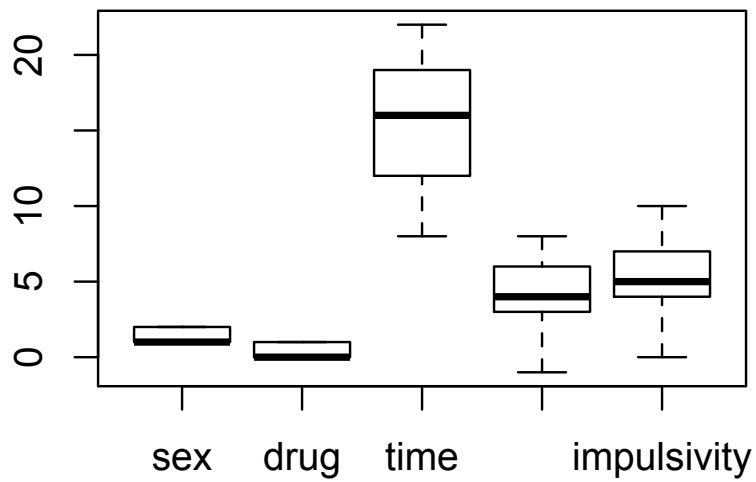
```
library(psych)                #Make this package active
#go to your browser, copy the results to the clipboard
sim.data <- read.clipboard()  #read the data from clipboard
describe(sim.data)            #basic descriptive statistics
```

```
            var   n  mean    sd median trimmed   mad min max range  skew kurtos  se
snum          1 100 50.50 29.01   50.5   50.50 37.06   1 100    99  0.00    -1.20 2.90
sex           2 100  1.43  0.50    1.0    1.41  0.00   1   2     1  0.28    -1.96 0.05
drug          3 100  0.46  0.50    0.0    0.45  0.00   0   1     1  0.16    -2.01 0.05
time          4 100 15.42  4.35   16.0   15.53  5.93   8  22    14 -0.09    -1.30 0.43
anxiety       5 100  4.10  1.98    4.0    4.17  1.48  -1   8     9 -0.36    -0.38 0.20
impulsivity   6 100  5.12  2.01    5.0    5.10  2.97   0  10    10  0.07    -0.61 0.20
arousal       7 100 63.02 36.49   64.0   63.15 32.62 -22 164   186  0.06     0.14 3.65
tension       8 100 23.57 13.36   21.5   22.39 12.60   4  61    57  0.71    -0.04 1.34
performance   9 100 63.79  9.96   65.0   63.94 10.38  35  88    53 -0.21    -0.06 1.00
cost         10 100  1.00  0.00    1.0    1.00  0.00   1   1     0   NaN      NaN 0.00
```

Simple descriptive graphics

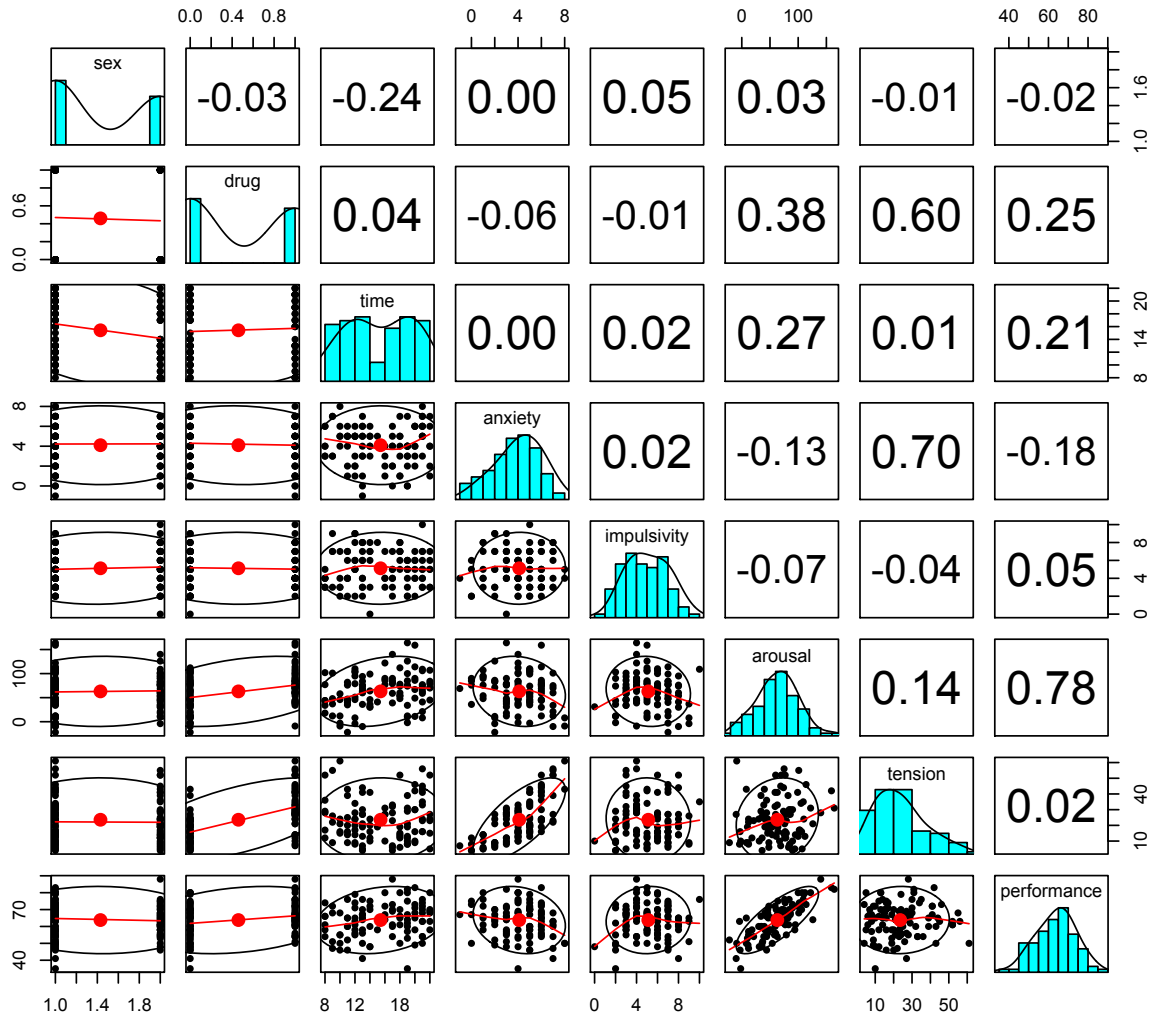
boxplot(sim.data) #this includes all variables
boxplot(sim.data[2:6]) #this just includes the important variales





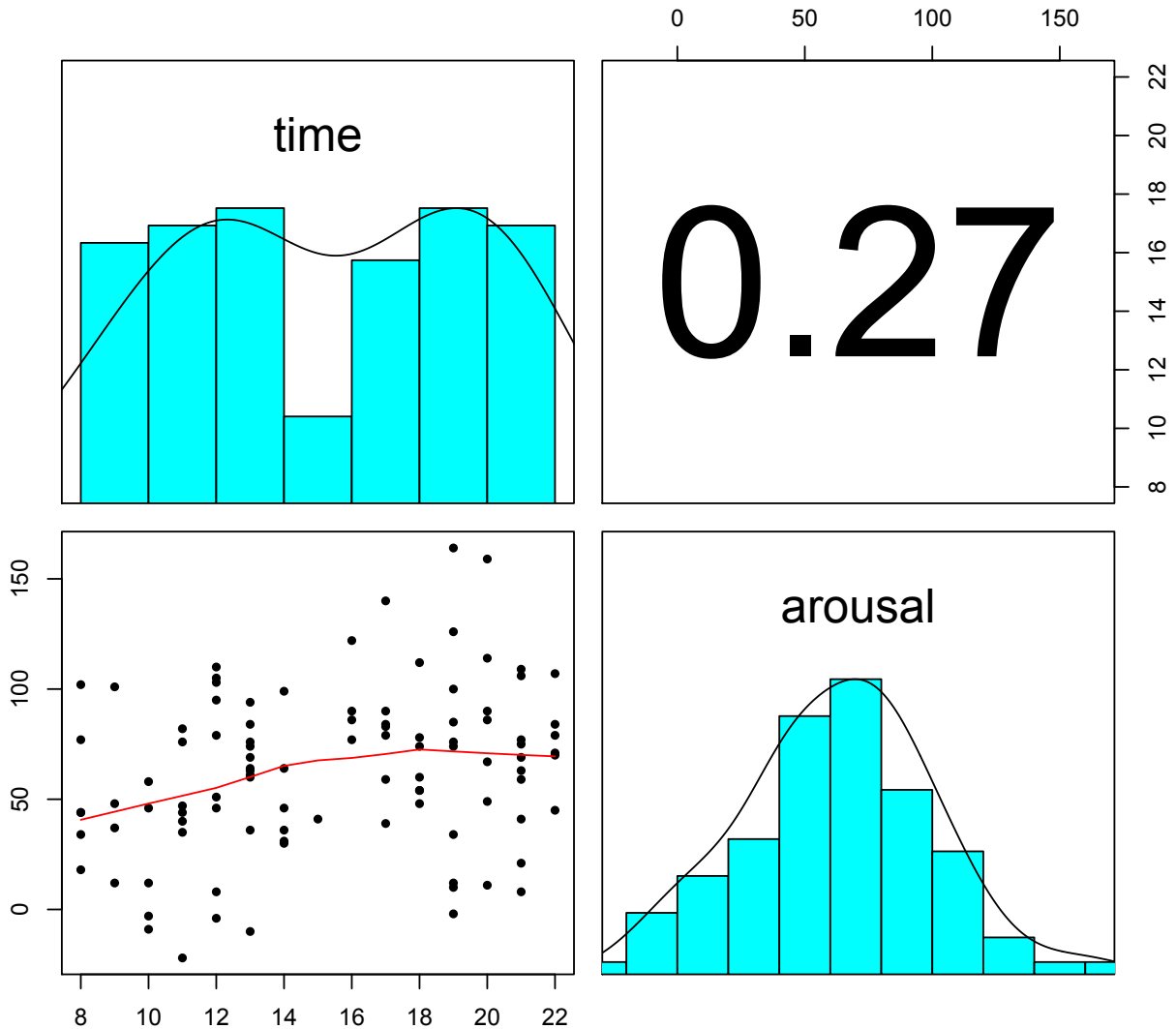
Graphical Summary: a SPLOM (scatter plot matrix)

```
op <- par(mfrow=c(1,1)) #change the figure back to a one panel

pairs.panels(sim.data[2:9]) #scatter plot matrix of variables 2-9
```



A more detailed plot of just two variables, showing the LOESS fit

```
pairs.panels(sim.data[c(4,7)],ellipses=FALSE)
```



Linear regression: Dependent Variables as a function of Independent Variables

1. We want to model the observed data (the DV) in terms of a set of predictors
   - Does the DV change as function of each predictor?
   - Does the effect of one predictor (IV) change as a function of some other predictor (IV)?
2. The basic linear model is that $y = x_1 + x_2 + x_1 * x_2 + \mathit{residual}$
   - The x's may be either categorical variables or continuous variables
   - If both are categorical, we call this an ANOVA
   - if one is categorical and the other continuous, we call it a moderated multiple regression
   - if both are continuous and we ignore interactions we call it multiple regression
   - if both are continuous and we consider interactions, it is a moderated multiple regression

Using the linear regression model requires centering the data

1. ANOVA and regression are equivalent models if
   - the Independent Variables are centered around 0.
   - Then the product terms (interactions) are not confounded with main effects.
2. Centering is done using the scale function
   - we set the scale parameter=FALSE to just center, not standardize.
   - we set the scale parameter=TRUE if we want standard scores.
3. But we need to turn the data back into a data frame when we are finished.

Centering the data for regression analysis

```
#dont type in the leading >      This is just a symbol that is used as a prompt
> cen.data <- scale(sim.data,scale=FALSE) #uses the scale function
> cen.data <- data.frame(cen.data) #convert to a data frame
> describe(cen.data)

            var   n mean    sd median trimmed   mad    min    max range   se
snum          1 100    0 29.01   0.00    0.00 37.06 -49.50  49.50    99 2.90
sex           2 100    0  0.50  -0.43   -0.02  0.00  -0.43   0.57     1 0.05
drug          3 100    0  0.50  -0.46   -0.01  0.00  -0.46   0.54     1 0.05
time          4 100    0  4.35   0.58    0.11  5.93  -7.42   6.58    14 0.43
anxiety       5 100    0  1.98  -0.10    0.08  1.48  -5.10   3.90     9 0.20
impulsivity   6 100    0  2.01  -0.12   -0.02  2.97  -5.12   4.88    10 0.20
arousal       7 100    0 36.49   0.98    0.13 32.62 -85.02 100.98   186 3.65
tension       8 100    0 13.36  -2.07   -1.18 12.60 -19.57  37.43    57 1.34
performance   9 100    0  9.96   1.21    0.15 10.38 -28.79  24.21    53 1.00
cost         10 100    0  0.00   0.00    0.00  0.00   0.00   0.00     0 0.00
>
```

Linear modeling (aka linear regression) is a generalization of ANOVA

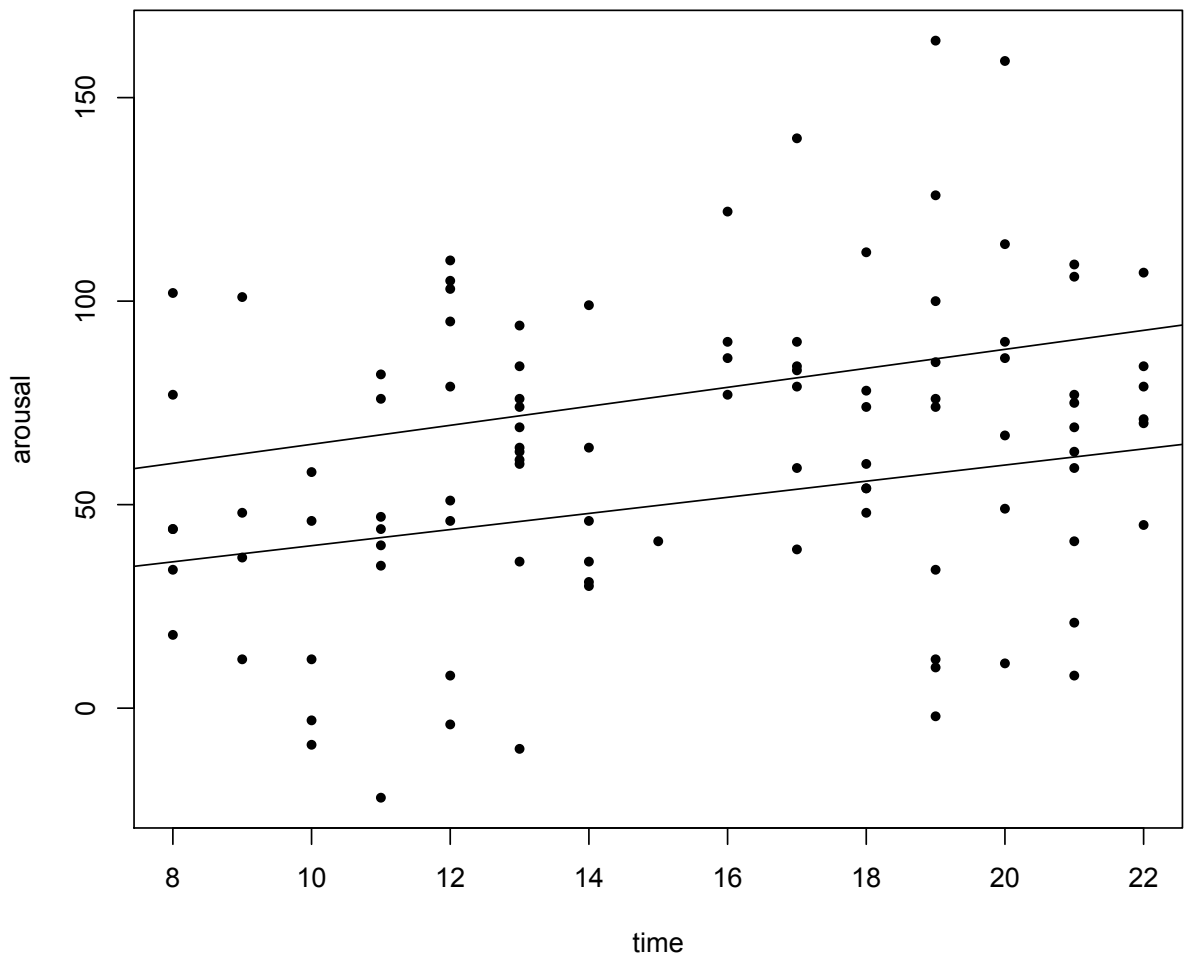
```
#specify the model
> model1 <- lm(arousal~drug * time,data=cen.data)
> summary(model1)   #now show the results

Call:
lm(formula = arousal ~ drug * time, data = cen.data)

Residuals:
    Min      1Q  Median      3Q     Max
-73.819 -22.319  -1.291  18.306  78.181

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.03407    3.29566  -0.010 0.991774
drug        26.81406    6.61261   4.055 0.000102 ***
time         2.14321    0.76361   2.807 0.006062 **
drug:time    0.35193    1.52333   0.231 0.817788
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1

Residual standard error: 32.92 on 96 degrees of freedom
Multiple R-squared: 0.2105, Adjusted R-squared: 0.1858
F-statistic: 8.533 on 3 and 96 DF,  p-value: 4.43e-05
```

Why graphics? Powerful graphics are valuable ways of showing results.

1. The previous analysis shows that drug has a positive effect on arousal (the slope was 26.8) as does time of day (the slope is 2.1)
   - But it would help if we could see this effect.
   - Draw it with a continuous independent variable (e.g., time of day) as the X axis, the Dependent variable as the Y axis, and a different line for the two drug conditions.
2. This is a two step process. First draw the data
3. Then draw the slopes.
   - arousal increases 26.8 units for one unit change in drug
   - arousal increases 2.1 units for one unit change in time (but there are more time units!)

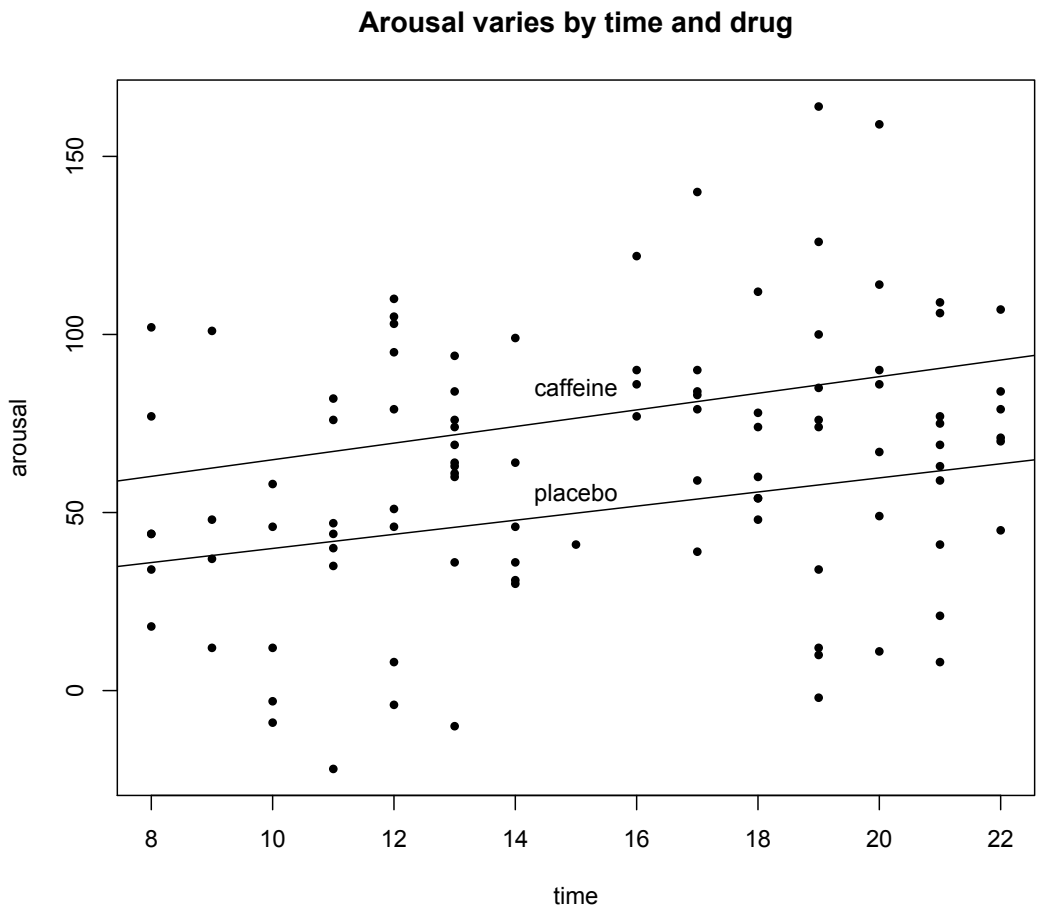
Graphing two experimental levels – the basic graphic

```
with(sim.data,plot(arousal~time))  #plot the data points
by(sim.data,sim.data$drug,function(x) abline(lm(arousal~time,data=x))) #add the lines
```



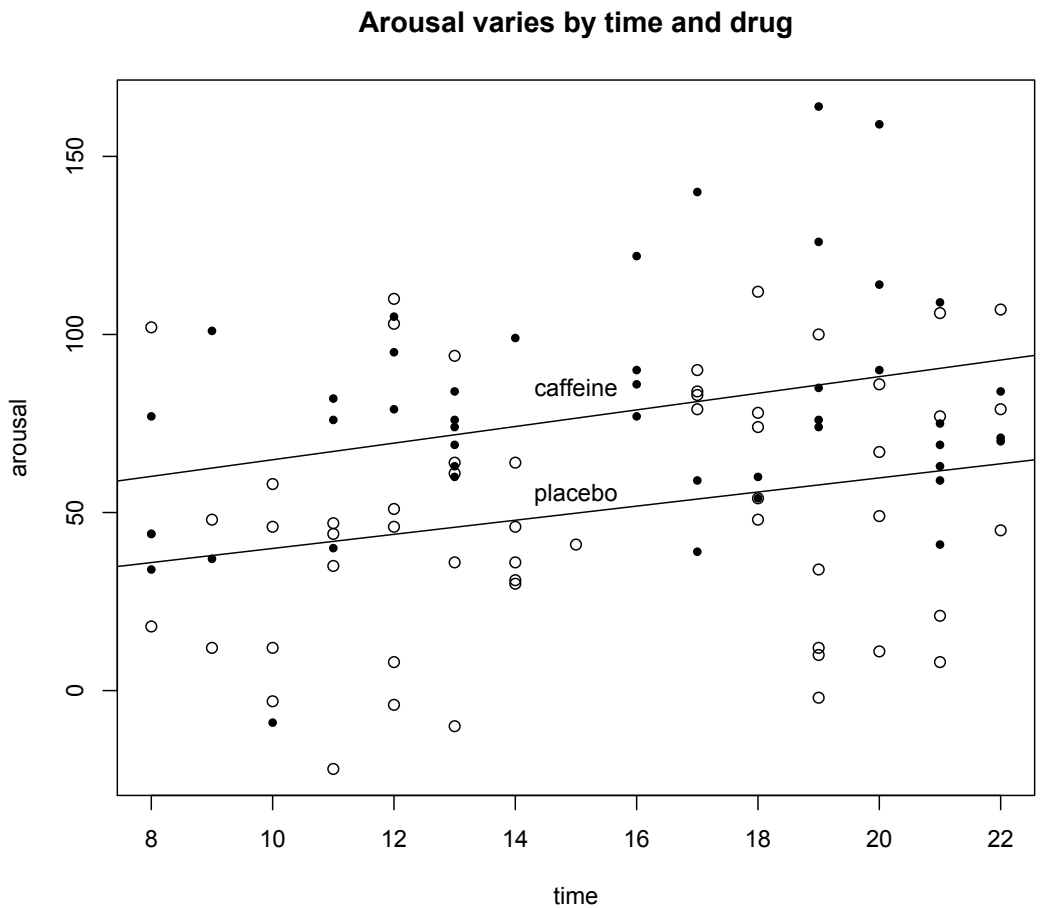
Graphing two experimental levels: Spice it up

```
with(sim.data,plot(arousal~time))  #plot the data points
by(sim.data,sim.data$drug,function(x) abline(lm(arousal~time,data=x))) #add the lines
 text(15,85,"caffeine")
 text(15,55,"placebo")
 title("Arousal varies by time and drug")
```



Graphing two experimental levels: Even more advanced graphing

```
with(sim.data,plot(arousal~time,pch=(21-drug)))  #plot the data points with different symbols
by(sim.data,sim.data$drug,function(x) abline(lm(arousal~time,data=x))) #add the lines
 text(15,85,"caffeine")
 text(15,55,"placebo")
 title("Arousal varies by time and drug")
```

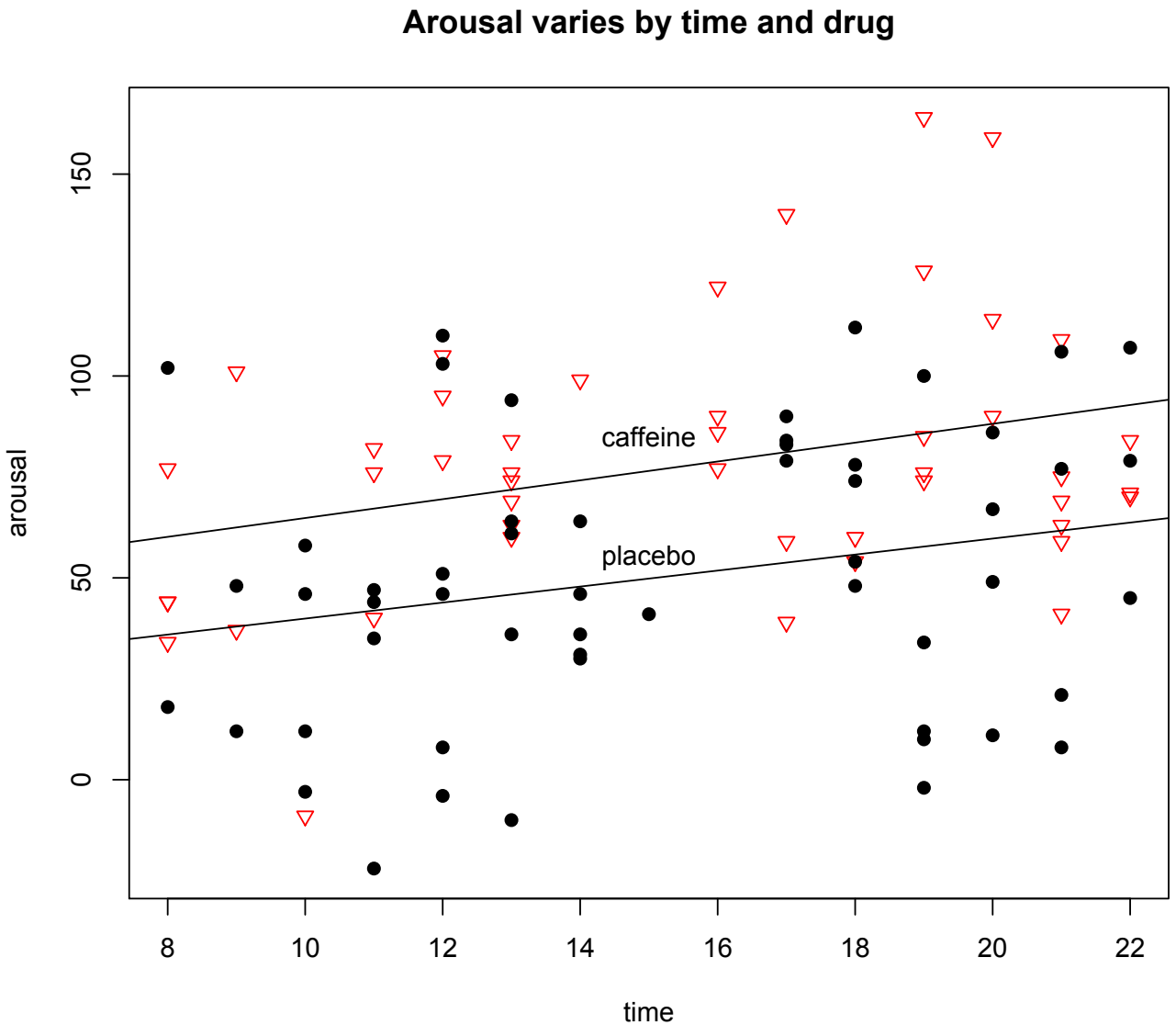


Even fancier graphics

Although you should not use color in publications, for slide shows it helps.

```
symb=c(19,25,3,23)  #choose some nice plotting symbols
colors=c("black","red","green","blue") #choose some nice colors
with(sim.data,plot(arousal~time,pch=symb[drug+1],
      col=colors[drug+1]) ) #plot the data points with different symbols
by(sim.data,sim.data$drug,function(x)
          abline(lm(arousal~time,data=x))) #add the lines
 text(15,85,"caffeine")
 text(15,55,"placebo")
 title("Arousal varies by time and drug")
```

The same graph but with colors



Anova as a special case of linear regression

- If variables are categorical, we frequently will treat the data using an Analysis of Variance (ANOVA)
- $Y = bX + c$ is the linear model
- ANOVA is a special case of the linear model where the predictor variables are categorical "factors" rather than continuous variables.
- We can specify variables as factors or use anova.

Compare ANOVA and lm

```
> mod0 <- aov(arousal~ drug * sex,data=cen.data)
> summary(mod0)
             Df Sum Sq Mean Sq F value    Pr(>F)
drug          1  19005 19004.8 16.2065 0.0001135 ***
sex           1    222   221.9  0.1892 0.6645207
drug:sex      1      5     5.4  0.0046 0.9458353
Residuals    96 112576  1172.7
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1
```

Subtle difference

```
>  mod2 <- lm(arousal ~drug *sex,data=cen.data)
> summary(mod2)

Call:
lm(formula = arousal ~ drug * sex, data = cen.data)

Residuals:
    Min      1Q  Median      3Q     Max
-85.926 -19.868  -2.594  22.493  87.074

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.007383   3.426134  -0.002 0.998285
drug        27.752247   6.874395   4.037 0.000109 ***
sex          3.006258   6.920706   0.434 0.664983
drug:sex    -0.946540  13.896081  -0.068 0.945835
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1

Residual standard error: 34.24 on 96 degrees of freedom
Multiple R-squared: 0.1459, Adjusted R-squared: 0.1192
F-statistic: 5.467 on 3 and 96 DF,  p-value: 0.001639
```

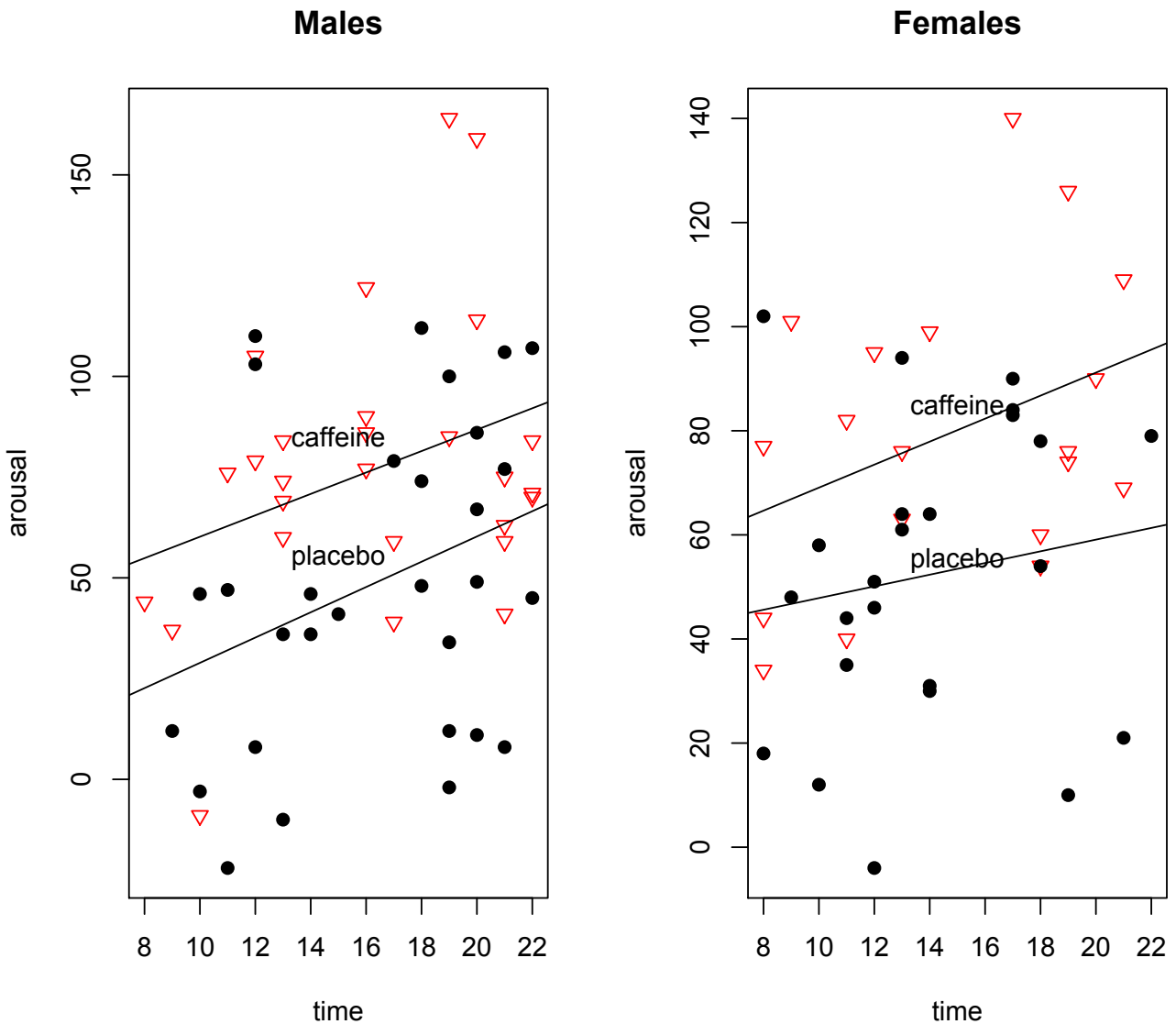
High level interactions can be decomposed into "simpler" effects

- Interpreting a high level (3 way or more) interaction is made easier if we consider lower level effects.
  - This can be done by subsetting the data into groups, and then doing the analysis within groups
  - Multiple ways this can be done (e.g., by sex, by drug, by level of something else)
- Most useful for doing multipanel graphics, but also for doing lower level regressions/anovas

Using the subset command to form two new data.frames

```
op <- par(mfrow=c(1,2)) #get ready for a two panel graph

> males <- subset(sim.data,sex=="1")   #form a new data.frame for the males
> dim(males)    #how many males and how many variables
[1] 57 10
> females <- subset(sim.data,sex=="2") #and another for the females
> dim(females)  #how many females and how many variables
[1] 43 10

symb=c(19,25,3,23)  #choose some nice plotting symbols
colors=c("black","red","green","blue") #choose some nice colors
 #plot the data points with different symbols
with(males,plot(arousal~time,pch=symb[drug+1],col=colors[drug+1]) )
 #add the lines
by(males,males$drug,function(x) abline(lm(arousal~time,data=x)))
 text(15,85,"caffeine")
 text(15,55,"placebo")
  title("Males")
 #plot the data points with different symbols
with(females,plot(arousal~time,pch=symb[drug+1],col=colors[drug+1]) )
by(females,females$drug,function(x) abline(lm(arousal~time,data=x))) #add the lines
 text(15,85,"caffeine")
 text(15,55,"placebo")
 title("Females")
```

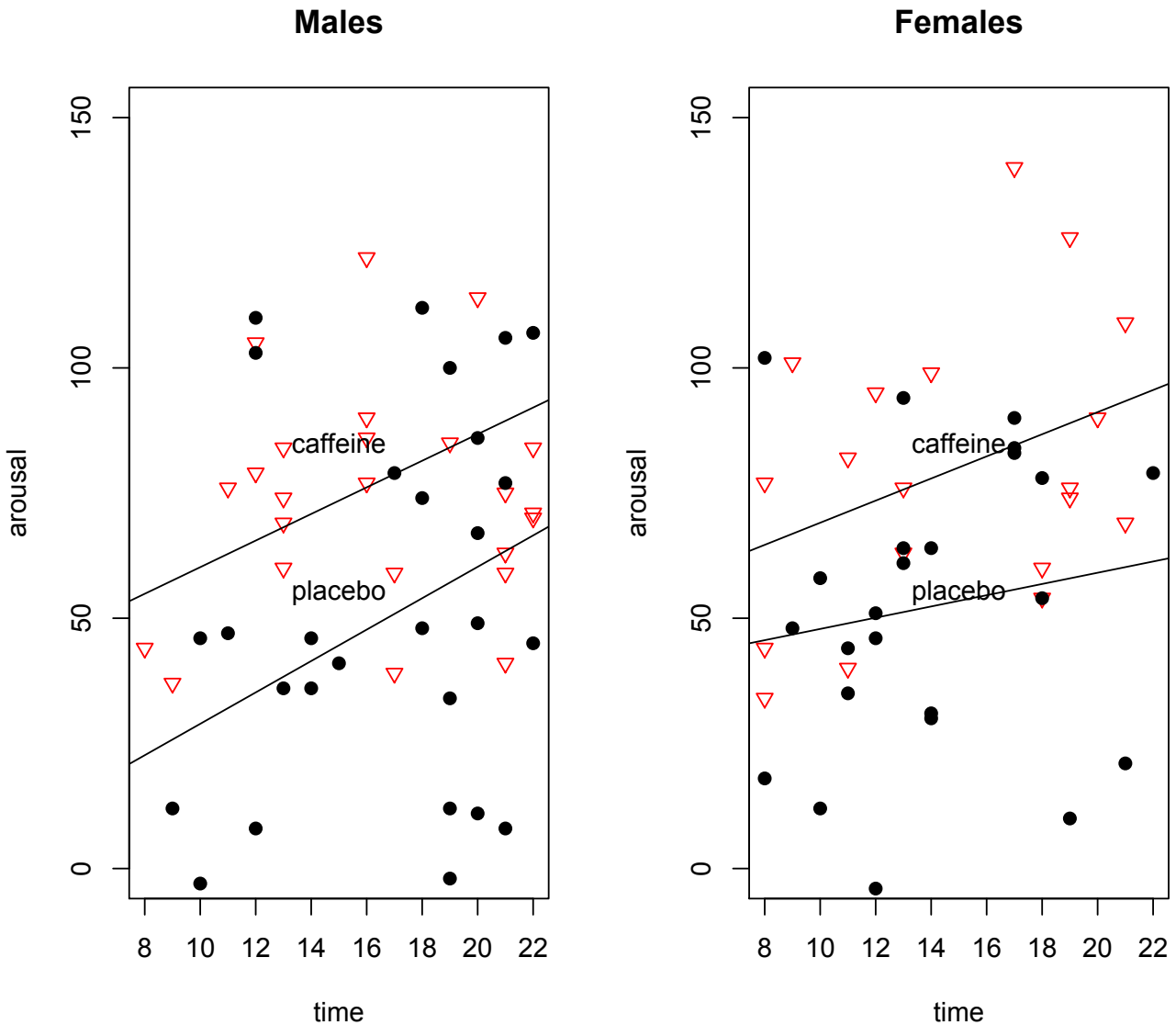
A two panel graph shows multiple effects - note the scales are different



modify that plot to use the same scale

```
op <- par(mfrow=c(1,2)) #get ready for a two panel graph

symb=c(19,25,3,23)  #choose some nice plotting symbols
colors=c("black","red","green","blue") #choose some nice colors
with(males,plot(arousal~time,pch=symb[drug+1],col=colors[drug+1],
                ylim=c(0,150)) ) #plot the data points with different symbols
by(males,males$drug,function(x) abline(lm(arousal~time,data=x))) #add the lines
 text(15,85,"caffeine")
 text(15,55,"placebo")
  title("Males")

with(females,plot(arousal~time,pch=symb[drug+1],col=colors[drug+1],
                  ylim=c(0,150)) ) #plot the data points with different symbols
by(females,females$drug,function(x) abline(lm(arousal~time,data=x))) #add the l
 text(15,85,"caffeine")
 text(15,55,"placebo")
 title("Females")

op <- par(mfrow=c(1,1)) #set it back to do one panel graphs
```

A two panel graph shows multiple effects – use the same scale



The `scrub` function allows you to recode your data

?scrub #ask what it does. The help (some function) command is most useful

```
A utility for basic data cleaning and recoding. Changes values outside of
minimum and maximum limits to NA.

Description

A tedious part of data analysis is addressing the problem of miscoded data that
need to be converted to NA. For a given data.frame or matrix, scrub will set
all values of columns from=from to to=to that are less than a set (vector) of
min values or more than a vector of max values to NA. Can also be used to
do basic recoding of data for all values=isvalue to newvalue.

Usage

scrub(x, where, min, max,isvalue,newvalue)
Arguments

x  a data frame or matrix
where  The variables to examine. (Can be by name or by column number)
min  a vector of minimum values that are acceptable
max  a vector of maximum values that are acceptable
isvalue  a vector of values to be converted to newvalue (one per variable)
newvalue  a vector of values to replace those that match isvalue
```

Recoding using the `scrub` function

```
new.data <- scrub(sim.data,where=4,min=15,newvalue=1)
new.data <- scrub(new.data,where=4,max=2,newvalue=2)
describe(new.data)
```

```
             var   n  mean    sd median trimmed   mad min max range  skew kurtosis   se
snum           1 100 50.50 29.01   50.5   50.50 37.06   1 100    99  0.00    -1.20 2.90
sex            2 100  1.43  0.50    1.0    1.41  0.00   1   2     1  0.28    -1.96 0.05
drug           3 100  0.46  0.50    0.0    0.45  0.00   0   1     1  0.16    -2.01 0.05
time           4 100  1.52  0.50    2.0    1.52  0.00   1   2     1 -0.08    -2.03 0.05
anxiety        5 100  4.10  1.98    4.0    4.17  1.48  -1   8     9 -0.36    -0.38 0.20
impulsivity    6 100  5.12  2.01    5.0    5.10  2.97   0  10    10  0.07    -0.61 0.20
arousal        7 100 63.02 36.49   64.0   63.15 32.62 -22 164   186  0.06     0.14 3.65
tension        8 100 23.57 13.36   21.5   22.39 12.60   4  61    57  0.71    -0.04 1.34
performance    9 100 63.79  9.96   65.0   63.94 10.38  35  88    53 -0.21    -0.06 1.00
cost          10 100  1.00  0.00    1.0    1.00  0.00   1   1     0   NaN      NaN 0.00
```

Use the recoded data in an ANOVA

```
>  mod4 <- aov(arousal~drug*time,data=new.data)
> summary(mod4)
print(model.tables(mod4,"means"),digits=3)
#report the means and the number of subjects/cell

             Df Sum Sq Mean Sq F value    Pr(>F)
drug          1  19005 19004.8 17.5944 6.096e-05 ***
time          1   9081  9081.0  8.4071  0.004633 **
drug:time     1     26    26.5  0.0245  0.875974
Residuals    96 103696  1080.2
---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1
Tables of means
Grand mean
63.02
 drug
drug
   0    1
50.3 78.0
 time
time
   1    2
53.1 72.1
```

Or, to treat them as factors, as we should

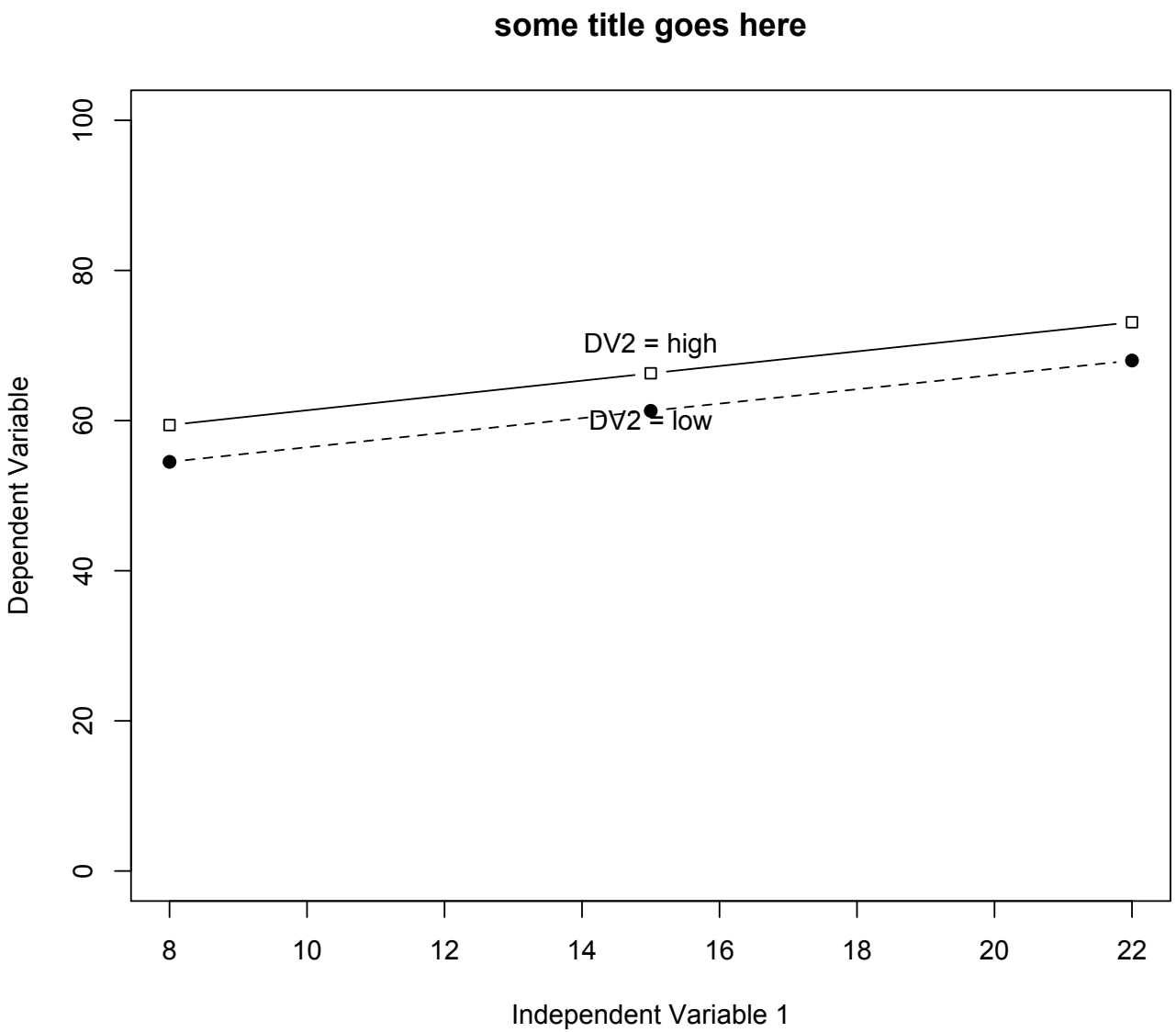
```
>  mod4 <- aov(arousal~as.factor(drug)*as.factor(time),data=new.data)
> summary(mod4)
> print(model.tables(mod4,"means"),digits=3)

                               Df Sum Sq Mean Sq F value    Pr(>F)
as.factor(drug)                 1  19005 19004.8 17.5944 6.096e-05 ***
as.factor(time)                 1   9081  9081.0  8.4071  0.004633 **
as.factor(drug):as.factor(time) 1     26    26.5  0.0245  0.875974
Residuals                      96 103696  1080.2                     ---
Signif. codes:  0 Ô***Õ 0.001 Ô**Õ 0.01 Ô*Õ 0.05 Ô.Õ 0.1 Ô Õ 1
 as.factor(drug)
       0  1
    50.3 78
rep 54.0 46      <- this is the number of subjects in the conditions
 as.factor(time)
       1    2
    53.1 72.1
rep 48.0 52.0
 as.factor(drug):as.factor(time)
              as.factor(time)
as.factor(drug) 1    2
            0   41.5 59.7
            rep 28.0 26.0
            1   66.5 86.8
            rep 20.0 26.0
```

Graphing anova results

Hard coding to show how to make a generic graph
Change the values to plot your data

```
IV1 <- c(8,15,22)    #the values of the first Independent Variable (time of day)

iv2low=c(54.5, 61.3 ,68.0)  #the Dependent variable values for low, medium and h
iv2high=c(59.4, 66.3, 73.1) #the Dependent variable values for low, mediuam and
IV1 <- c(8,15,22)    #the values of the first Independent Variable

plot(IV1,iv2low,xlim=c(8,22),ylim=c(0,100),type="b",lty="dashed",
xlab="Independent Variable 1" ,
ylab="Dependent Variable",
 pch=19,main="some title goes here")    #plot character is closed circle

points(IV1,iv2high,pch=22,type="b") #plot character is an open square
text(15,60,"DV2 = low")    #change the location of these labels to fit the data
text(15,70,"DV2 = high")
```

A generic plot



For more guidance

- Katherine Funkhouser's guide; http://personality-project.org/revelle/syllabi/205/analysing-data.pdf
- The 205 guide http://personality-project.org/r/r.205.tutorial.html
- A longer guide: http://personality-project.org/r/r.guide.html
- The guide to the simulation experiment: http://personality-project.org/revelle/syllabi/205/simulation/simulating-experiments.pdf

Don't Panic