Psychology 350: An introduction to R for Psychological Research
Week 5b: The linear model

William Revelle

Department of Psychology
Northwestern University
Evanston, Illinois USA

April, 2024

Outline

Causal effects

1. The experimentalist wants to know how much *changing* one variable (X) produces changes in another (Y). Typically we call X and Y the Independent Variable and the Dependent Variable.
2. This leads to an experimental manipulation of X into two levels (0 and 1) and then the observation of the values of Y for those two conditions and their expectations are
3. $\mathbb{E}(Y|X=0) = \bar{Y}_0$ and $\mathbb{E}(Y|X=1) = \bar{Y}_1$
4. Find the means for these two and take their difference : $D = \bar{Y}_1 - \bar{Y}_0$
5. But these means reflect variability and scale (kg vs. gms). So find
6. $d = \frac{\bar{Y}_1 - \bar{Y}_0}{sd}$ as a measure of the effect size of X

t: the mean difference in comparison to the standard error

1. Gossett/Student (1908) expressed the mean difference in terms of the standard error of the difference
2. se of difference is twice the square root of the pooled within group squared standard errors:
3. $$se_d = \sqrt{\frac{sd_0^2}{n_0 - 1} + \frac{sd_1^2}{n_1 - 1}}$$
4. $$t = \frac{\bar{Y}_1 - \bar{Y}_0}{\sqrt{\frac{sd_0^2}{n_0 - 1} + \frac{sd_1^2}{n_1 - 1}}}$$
5. Gossett/Student derived the distribution of this statistic for small samples.
6. Therefore, t varies as the effect size and the sample size: $t = \frac{d\sqrt{df}}{2}$

Effect size

1. There are many ways of reporting how two groups differ. Cohen's d statistic (Cohen, 1988) is just the differences of means expressed in terms of the pooled within group standard deviation. This is insensitive to sample size.
2. r is a universal measure of effect size that is a simple function of d, but is bounded -1 to 1.
3. The t statistic is merely d * sqrt(df)/2 and thus reflects sample size.
4. Confidence intervals for Cohen's d may be found by converting the d to a t, finding the confidence intervals for t, and then converting those back to ds. This take advantage of the uniroot function and the non-centrality parameter of the t distribution.
5. See `cohen.d`

cohen.d on the sat.act data set

R code

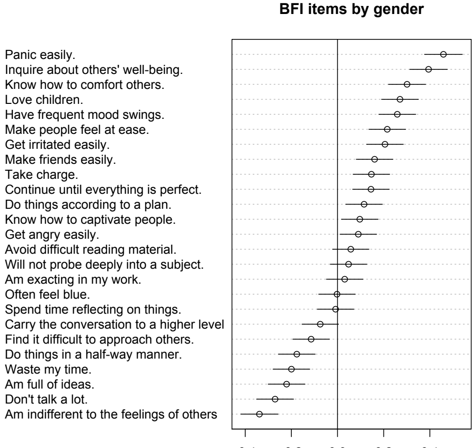
```
cohen.d(sat.act,"gender")  #t test on all the subjects
cohen.d(sat.act[1:40,],"gender")  #and then just the first 40 subjects
```

```
 cohen.d(sat.act,"gender")  #
Call: cohen.d(x = sat.act, group = "gender")
Cohen d statistic of difference between two means
          lower effect upper
education  0.03   0.18  0.34
age       -0.20  -0.04  0.11
ACT       -0.23  -0.08  0.08
SATV      -0.19  -0.04  0.12
SATQ      -0.51  -0.35 -0.19
Multivariate (Mahalanobis) distance between groups
[1] 0.52
r equivalent of difference between two means
education       age       ACT      SATV      SATQ
     0.09     -0.02     -0.04     -0.02     -0.17
Call: cohen.d(x = sat.act[1:40, ], group = "gender")
Cohen d statistic of difference between two means
          lower effect upper
education -0.15   0.49  1.12
age       -0.60   0.03  0.65
ACT       -0.49   0.14  0.76
SATV      -0.61   0.02  0.64
SATQ      -1.10  -0.47  0.17
Multivariate (Mahalanobis) distance between groups
[1] 0.82
r equivalent of difference between two means
education       age       ACT      SATV      SATQ
     0.24      0.01      0.07      0.01     -0.23
```

Plotting cohen.d for bfi items by gender

R code

```
cd <- cohen.d(bfi[1:26],"gender",
      dictionary=bfi.dictionary[,2,drop=FALSE])
error.dots(cd,head=13,tail=13,main="BFI items by gender")
abline(v=0)
```



The t-test and effect size

1. The t-test is an effect size/standard error ($\sigma_{\bar{x}}$) of effect size. (For equal size groups)

$$es = \frac{x_1 - x_2}{\sqrt{(\sigma^2_{x_1} + \sigma^2_{x_2})/2}} \tag{1}$$

and

$$\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2_x}{df}} \tag{2}$$

$$t = es\frac{\sqrt{df}}{2} \tag{3}$$

2. If expressed as a regression, slope reflects how much y changes for a unit change in x.
3. Note how effect size is not affected by sample size, t is.

t.test is sensitive to sample size

R code

```
t.test(education ~ gender,data=sat.act)
t.test(education ~ gender,data=sat.act[1:40,])
```

```
        Welch Two Sample t-test

data:  education by gender
t = -2.2299, df = 453.96, p-value = 0.02624
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.48935928 -0.03087916
sample estimates:
mean in group 1 mean in group 2
       2.995951        3.256071

> t.test(education ~ gender,data=sat.act[1:40,])

        Welch Two Sample t-test

data:  education by gender
t = -1.4356, df = 28.257, p-value = 0.1621
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.3601898  0.2389777
sample estimates:
mean in group 1 mean in group 2
       3.166667        3.727273
```

More on effect size

1. In a recent paper with Alice Eagly,(Eagly and Revelle, 2022) we showed how effect sizes can vary by aggregating items.
2. At the item level, there are many very small gender differences, but when pooled into scales, the differences are quite noticeable.
3. We made us of the Mahalanobis (1936) distance. (See McLachlan (1999) for a discussion of the M distance, and Del Giudice (2009); Del Giudice et al. (2012) for applications.)
4. M distance is just the distance in multivariate space between two centroids. It is $\sqrt{\mathbf{dR}^{-1}\mathbf{d}'}$. where **d** is a vector of distances and **R** is the correlation matrix.
5. Reported by `cohen.d`.

The `athenstaedt` data set

1. Included in *psychTools* is a dataset taken from Ursala Athenstaedt (2003)
2. Ursala Athenstaedt (2003) reported several analyses of items and scales measuring Gender Role Self-Concept.
3. Eagly and Revelle (2022) have used these data in an analysis of the power of aggregation.
4. Here are the original items as well as the three scales Eagly and Revelle (2022).
5. The accompanying Athenstaedt.dictionary may be used to see the items.

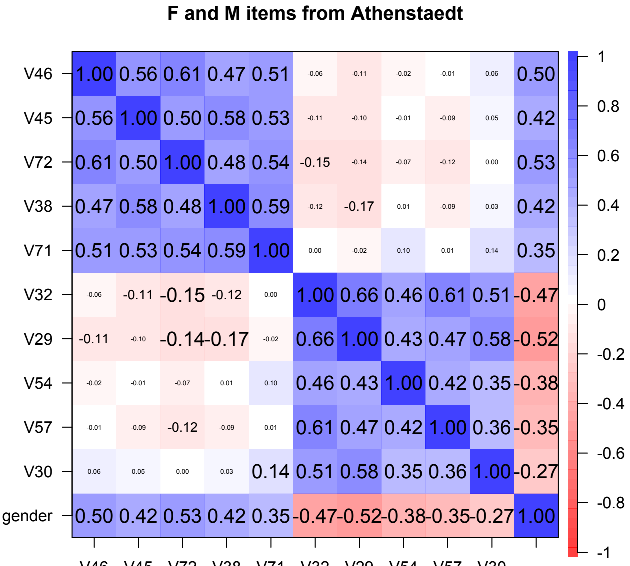
Show some of the items

R code

```
lookupFromKeys(Athenstaedt.keys[7:8],
      dictionary=Athenstaedt.dictionary)
```

```
$F5
    ItemLabel                Item
V46       V46      Sew on a Button
V45       V45   Change Bed Sheets
V72       V72      Do the Ironing
V38       V38 Dust the Furniture
V71       V71        Wash Windows

$M5
    ItemLabel                      Item
V32       V32           Do Repair Work
V29       V29             Change Fuses
V54       V54              Shovel Snow
V57       V57 Do Home Improvement Jobs
V30       V30            Clean a Drain
```

The items in these scales correlate within but not between scales



Scoring the Athenstaedt items

R code

```
scales<- scoreOverlap(Athenstaedt.keys,Athenstaedt)
scatterHist(Femininity ~ Masculinity + gender, data =Athenstaedt,
cex.point=.4,smooth=FALSE, correl=FALSE,d.arrow=TRUE,col=c("blue","re
   lwd=4,  cex.main=1.5,main="Scatter Plot and Density",cex.axis=2)
```

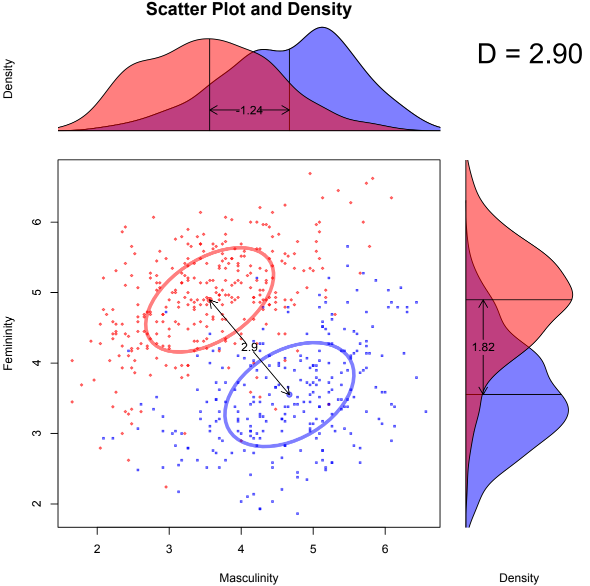
```
Scale intercorrelations corrected for item overlap and attenuation
 adjusted for overlap correlations below the diagonal, alpha on the diagonal
 corrected correlations above the diagonal:
```

	Femininity	Masculinity	MF	F10	M10	MF20	F5	M5	MF10
Femininity	0.900	-0.090	0.81	0.931	-0.141	0.75	0.885	-0.159	0.75
Masculinity	-0.079	0.875	-0.66	-0.082	0.976	-0.71	-0.050	0.961	-0.70
MF	0.719	-0.580	0.88	0.749	-0.684	0.99	0.695	-0.690	0.98
F10	0.831	-0.072	0.66	0.886	-0.092	0.75	0.987	-0.113	0.78
M10	-0.125	0.852	-0.60	-0.081	0.871	-0.73	-0.056	0.995	-0.72
MF20	0.652	-0.614	0.85	0.648	-0.624	0.85	0.714	-0.737	1.02
F5	0.775	-0.044	0.60	0.858	-0.048	0.61	0.853	-0.077	0.74
M5	-0.137	0.817	-0.59	-0.096	0.843	-0.62	-0.065	0.824	-0.71
MF10	0.626	-0.573	0.81	0.644	-0.585	0.82	0.600	-0.569	0.77

Two separate domains: items and scales that correlate with being Male or Female, form reliable scales, but the scales are independent.

We can show this at the item level using the `scatterHist` function.

Analysis of the Athenstaedt data.



The linear model and its special cases

There are many forms of the linear model.

1. $\hat{y} = b_1 x + e$ is the classic regression model, where $b_1 = \frac{cov_{xy}}{var_x}$.
2. If x is a dichotomous variable, this is equivalent to a t-test or if there are more than two categories, as an Analysis of Variance.
3. The use of dichotomous variables is most frequently seen in experimental designs where we have two values of some experimental variable. We think of x causing y, and typically refer to x as an Independent Variables causing y, the Dependent Variable.
4. If expressed as a t-test, this is difference of means, divided by the standard error of the difference of means.
5. It is perhaps better to think of a t as an *effect size* divided by its standard error. The effect size is the difference in means divided by the pooled within group standard deviation:

Regression

1. Typical model is that X causes Y
   $\hat{y} = b_{x1}x + e$
2. The slope (b) is the ratio of the covariance of x and y divided by the variance of x.
   $b_{x1} = \frac{cov_{xy}}{var_x} = \frac{\sigma_{xy}}{\sigma_x^2}$
3. But, if we think of y causing x, this becomes:
4. Y causes X
   $\hat{x} = b_{y1}y + e$ and
   $b_{y1} = \frac{cov_{xy}}{var_t}$
5. If we are unsure of the direction of causality, we can find the geometric average of the two regressions and find
   $r_{xy} = \sqrt{b_{x1}b_{y1}} = \frac{\sigma_{xy}}{\sigma_y^2} = \sqrt{\frac{\sigma_{xy}}{\sigma_x^2}\frac{\sigma_{xy}}{\sigma_y^2}} = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2\sigma_y^2}} = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$
6. $\hat{y} = b_1x_1 + b_2x_2 + e$ Multiple regression. If x_1 and x_2 are categorical, this is also an analysis of variance.

Co-relationships (see week 3)

1. Descriptive measures of relationship
   - Do two (or more) variables co-vary?
2. Galton (1888) reported a method of measuring the "co-relation" of two measures
3. Pearson (1896) formalized this as the Pearson Product Moment Correlation Coefficient

$$\rho = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}}$$

where x and y are deviation scores from the mean

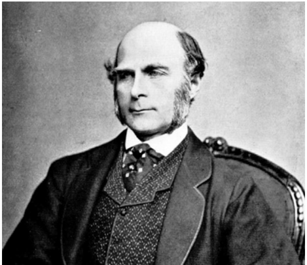
$$x = X - \bar{X} = X - \frac{\Sigma x}{N} \qquad y = Y - \bar{Y} = X - \frac{\Sigma x}{N}$$

4. Spearman (1904) expressed this in terms of rank orders.
5. in R we use the (cor) function

Francis Galton 1822-1911

Francis Galton (1822-1911) was among the most influential psychologists of the 19th century. He did pioneering work on the correlation coefficient, behavior genetics and the measurement of individual differences. He introspectively examined the question of free will and introduced the lexical hypothesis to the study of personality and character. In addition to psychology, he did pioneering work in meteorology and introduced the scientific use of fingerprints. Whenever he could, he counted.

http://personality-project.org/revelle/publications/galton.pdf



Karl Pearson 1857-1936

Carl (Karl) Pearson was among the most influential statisticians of the early 20th century. Founder of the statistics department at University College London. He developed the Pearson Product Moment Correlation Coefficient, its special case the ϕ coefficient, and the tetrachoric correlation. Major behavior geneticist and eugenicist.

Charles Spearman 1863-1945

Charles Spearman (1863-1945) was the leading psychometrician of the early 20th century. His work on the classical test theory, factor analysis, and the g theory of intelligence continues to influence psychometrics, statistics, and the study of intelligence. More than 100 years after their publication, his most influential papers remain two of the most frequently cited articles in psychometrics and intelligence. http://personality-project.org/revelle/publications/spearman.pdf

Galton's height data

Table: The relationship between the average of both parents (mid parent) and the height of their children. The basic data table is from Galton (1886) who used these data to introduce reversion to the mean (and thus, linear regression). The data are available as part of the **UsingR** or **psych** packages.

```
> library(psych)
> data(galton)
> galton.tab <- table(galton)
> galton.tab[order(rank(rownames(galton.tab)),decreasing=TRUE),] #sort it by decreasing row va
       child
parent 61.7 62.2 63.2 64.2 65.2 66.2 67.2 68.2 69.2 70.2 71.2 72.2 73.2 73.7
  73      0    0    0    0    0    0    0    0    0    0    0    1    3    0
  72.5    0    0    0    0    0    0    0    1    2    1    2    7    2    4
  71.5    0    0    0    0    1    3    4    3    5   10    4    9    2    2
  70.5    1    0    1    0    1    1    3   12   18   14    7    4    3    3
  69.5    0    0    1   16    4   17   27   20   33   25   20   11    4    5
  68.5    1    0    7   11   16   25   31   34   48   21   18    4    3    0
  67.5    0    3    5   14   15   36   38   28   38   19   11    4    0    0
  66.5    0    3    3    5    2   17   17   14   13    4    0    0    0    0
  65.5    1    0    9    5    7   11   11    7    7    5    2    1    0    0
  64.5    1    1    4    4    1    5    5    0    2    0    0    0    0    0
  64      1    0    2    4    1    2    2    1    1    0    0    0    0    0
```

Galton's height data

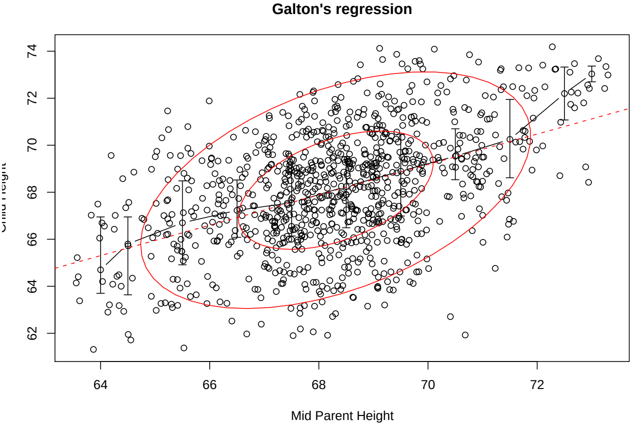


Figure: Galton's data can be plotted to show the relationships between mid parent and child heights. Because the original data are grouped, the data points have been *jittered* to emphasize the density of points along the median. The bars connect the first, 2nd (median) and third quartiles. The dashed line is the best fitting linear fit, the ellipses represent one and two standard deviations from the mean.

Bivariate Regression

X Y ϵ

$$\begin{tikzcd} \boxed{X} \arrow[r, "\beta_{y.x}"] & \boxed{Y} & \epsilon \arrow[l] \end{tikzcd}$$

$$y = \hat{y} + \epsilon = \beta_{y.x}x + \epsilon$$

$$\beta_{y.x} = \frac{\sigma_{xy}}{\sigma_x^2}$$

$$\epsilon = y - \hat{y}$$

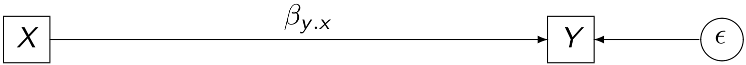
$$\sum(\epsilon^2) = \sum(y - \hat{y})^2 = \sum(y - \beta_{y.x}x)^2 = \sum(y^2 - 2y\beta_{y.x}x + (\beta_{y.x}x)^2)$$

Minimize $\sum(\epsilon^2) w.r.t. \beta => \frac{d(\epsilon^2)}{d\beta} = 0 => -2\sigma_{xy} + 2\beta_{y.x}\sigma_x^2 = 0 =>$

$$\beta_{y.x} = \frac{\sigma_{xy}}{\sigma_x^2}$$

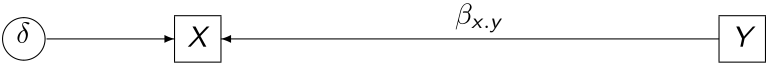
Bivariate Regression

δ X Y ϵ



$$y = \hat{y} + \epsilon = \beta_{y.x}x + \epsilon$$

$$\beta_{y.x} = \frac{\sigma_{xy}}{\sigma_x^2}$$

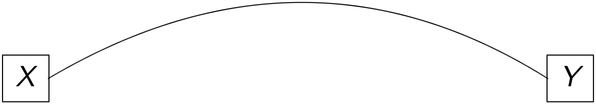


$$x = \hat{x} + \delta = \beta_{x.y}y + \delta$$

$$\beta_{y.x} = \frac{\sigma_{xy}}{\sigma_y^2}$$

Bivariate Correlation is the geometric average of the two regressions

X Y



$x = \hat{x} + \delta = \beta_{x.y} y + \delta$

$y = \hat{y} + \epsilon = \beta_{y.x} x + \epsilon$

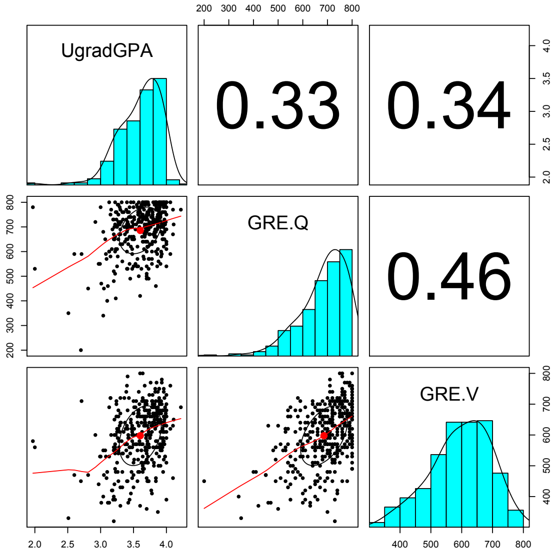
$\beta_{y.x} = \frac{\sigma_{xy}}{\sigma_y^2}$

$\beta_{y.x} = \frac{\sigma_{xy}}{\sigma_x^2}$

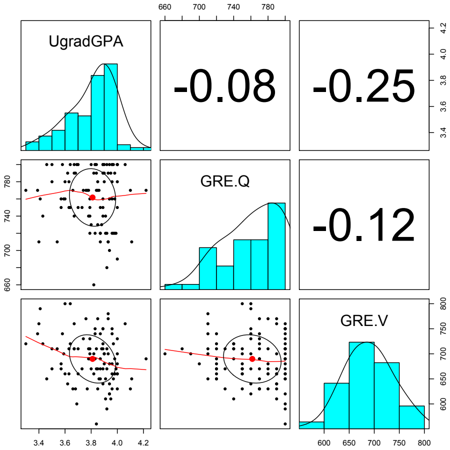
$$r_{xy} = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2 \sigma_y^2}}$$

$$r_{xy} = \sigma_{z_x z_y} \text{(the covariance of standard scores)}$$

Scatter Plot Matrix showing correlation and LOESS regression

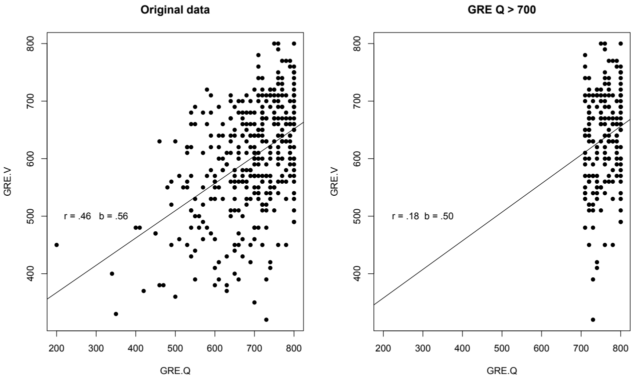


The effect of selection on the correlation



- Consider what happens if we select a subset
  - The "Oregon" model
  - (GPA + (V+Q)/200) > 11.6
- The range is truncated, but even more important, by using a compensatory selection model, we have changed the sign of the correlations.

Regression and restriction of range



Although the correlation is very sensitive, regression slopes are relatively insensitive to restriction of range.

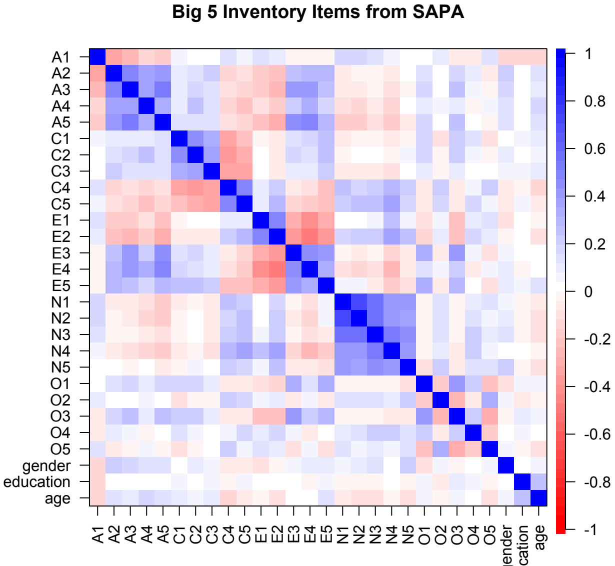
R code for regression figures

R code

```
 datafilename="http://personality-project.org/r/datasets/psychometrics
  mydata =read.table(datafilename,header=TRUE)  #read the data file
gradq <- subset(gradf,gradf[2]>700) #choose the subset
with(gradq,lm(GRE.V ~ GRE.Q)) #do the regression
```

```
Call:
lm(formula = GRE.V ~ GRE.Q)
Coefficients:
(Intercept)        GRE.Q
   258.1549       0.4977
#show the graphic
op <- par(mfrow=c(1,2))  #two panel graph
 with(gradf,{
 plot(GRE.V ~ GRE.Q,xlim=c(200,800),main='Original data', pch=16)
 abline(lm(GRE.V ~ GRE.Q))
 })
 text(300,500,'r = .46   b = .56')
 with(gradq,{
 plot(GRE.V ~ GRE.Q,xlim=c(200,800),main='GRE Q > 700',pch=16)
 abline(lm(GRE.V ~ GRE.Q))
 })
  text(300,500,'r = .18  b = .50')
 op <- par(mfrow=c(1,1)) #switch back to one panel
```

Show many correlations with a heat map using `cor.plot`.



Alternative versions of the correlation coefficient

Table: A number of correlations are Pearson r in different forms, or with particular assumptions. If $r = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}$, then depending upon the type of data being analyzed, a variety of correlations are found.

Coefficient	symbol	X	Y	Assumes
Pearson	r	continuous	continuous	
Spearman	rho (ρ)	ranks	ranks	
Point bi-serial	r_{pb}	dichotomous	continuous	
Phi	ϕ	dichotomous	dichotomous	
Bi-serial	r_{bis}	dichotomous	continuous	normality
Tetrachoric	r_{tet}	dichotomous	dichotomous	normality
Polychoric	r_{pc}	categorical	categorical	normality

use `cor` for the first 4, `biserial`, `tetrachoric`, `polychoric` to find these values.

The ϕ coefficient is just a Pearson r on dichotomous data

Table: The basic table for a phi, ϕ coefficient, expressed in raw frequencies in a four fold table is taken from Pearson and Heron (1913)

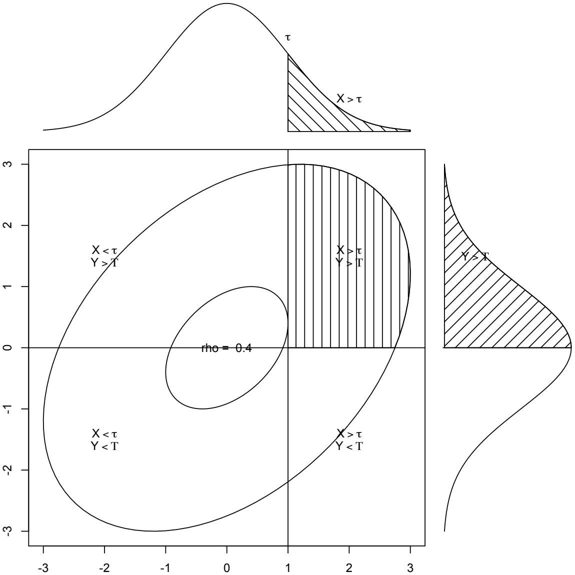
	Success	Failure	Total
Accept	A	B	R_1 = A + B
Reject	C	D	R_2 = C + D
Total	C_1 = A + C	C_2=B + D	n = A + B + C + D

In terms of the raw data coded 0 or 1, the *phi coefficient* can be derived directly by direct substitution, recognizing that the only non zero product is found in the A cell

$$n\sum X_iY_i - \sum X_i \sum Y_i = nA - R_1C_1$$

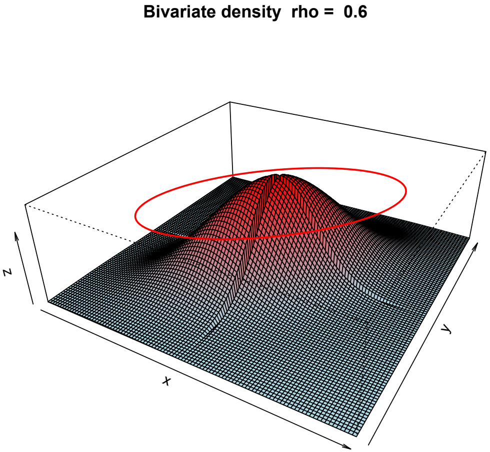
$$\phi = \frac{AD - BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}. \tag{4}$$

The tetrachoric correlation estimates the latent correlation



The tetrachoric correlation estimates the latent correlation

`tetrachoric` iteratively estimates the tetrachoric correlalation.



Cautions about correlations–The Anscombe data set

Consider the following 8 variables

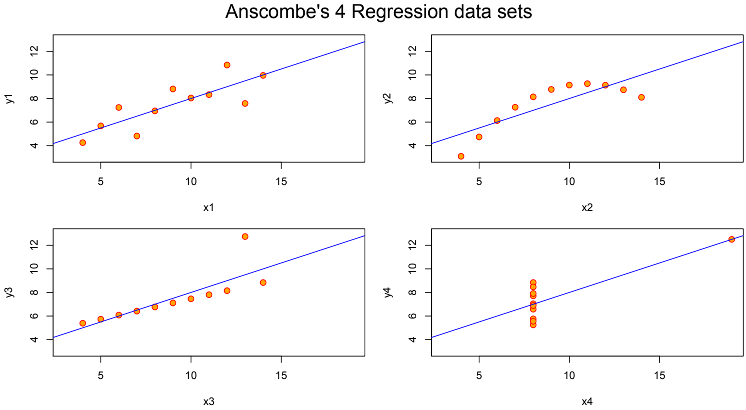
	var	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosi
x1	1	11	9.0	3.32	9.00	9.00	4.45	4.00	14.00	10.00	0.00	-1.2
x2	2	11	9.0	3.32	9.00	9.00	4.45	4.00	14.00	10.00	0.00	-1.2
x3	3	11	9.0	3.32	9.00	9.00	4.45	4.00	14.00	10.00	0.00	-1.2
x4	4	11	9.0	3.32	8.00	8.00	0.00	8.00	19.00	11.00	2.47	11.0
y1	5	11	7.5	2.03	7.58	7.49	1.82	4.26	10.84	6.58	-0.05	-0.5
y2	6	11	7.5	2.03	8.14	7.79	1.47	3.10	9.26	6.16	-0.98	0.8
y3	7	11	7.5	2.03	7.11	7.15	1.53	5.39	12.74	7.35	1.38	4.3
y4	8	11	7.5	2.03	7.04	7.20	1.90	5.25	12.50	7.25	1.12	3.1

Cautions, Anscombe continued

With regressions of

```
             Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 3.0000909  1.1247468 2.667348 0.025734051
x1          0.5000909  0.1179055 4.241455 0.002169629

[[2]]
            Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 3.000909  1.1253024 2.666758 0.025758941
x2          0.500000  0.1179637 4.238590 0.002178816

[[3]]
             Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 3.0024545  1.1244812 2.670080 0.025619109
x3          0.4997273  0.1178777 4.239372 0.002176305

[[4]]
             Estimate Std. Error  t value    Pr(>|t|)
(Intercept) 3.0017273  1.1239211 2.670763 0.025590425
x4          0.4999091  0.1178189 4.243028 0.002164602
```

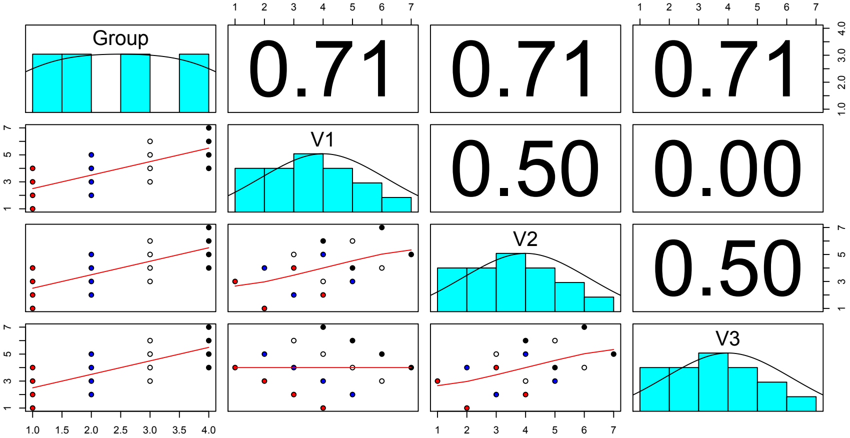
Cautions about correlations: Anscombe data set



Further cautions about correlations–the problem of levels

1. Correlations taken at one level of analysis can be unrelated to those at another level
2. $r_{xy} = \eta_{x_{wg}} * \eta_{y_{wg}} * r_{xy_{wg}} + \eta_{x_{bg}} * \eta_{y_{bg}} * r_{xy_{bg}}$
3. Where η is the correlation of the data with the within group values, or the group means.
4. The within group and between group correlations can even be of different sign!
5. The `withinBetween` data set is an example of this problem.
6. The `statsBy` function will find the within and between group correlations for this kind of multi-level design.

Cautions about correlations: Within versus between groups



The ubiquitous correlation coefficient

Table: Alternative Estimates of effect size. Using the correlation as a scale free estimate of effect size allows for combining experimental and correlational data in a metric that is directly interpretable as the effect of a standardized unit change in x leads to r change in standardized y.

Statistic	Estimate	r equivalent	as a function of r
Pearson correlation	$r_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y}$	r_{xy}	
Regression	$b_{y.x} = \frac{C_{xy}}{\sigma_x^2}$	$r = b_{y.x} \frac{\sigma_y}{\sigma_x}$	$b_{y.x} = r \frac{\sigma_x}{\sigma_y}$
Cohen's d	$d = \frac{X_1 - X_2}{\sigma_x}$	$r = \frac{d}{\sqrt{d^2+4}}$	$d = \frac{2r}{\sqrt{1-r^2}}$
Hedge's g	$g = \frac{X_1 - X_2}{s_x}$	$r = \frac{g}{\sqrt{g^2+4(df/N)}}$	$g = \frac{2r\sqrt{df/N}}{\sqrt{1-r^2}}$
t - test	$t = \frac{d\sqrt{df}}{2}$	$r = \sqrt{t^2/(t^2+df)}$	$t = \sqrt{\frac{r^2 df}{1-r^2}}$
F-test	$F = \frac{d^2 df}{4}$	$r = \sqrt{F/(F+df)}$	$F = \frac{r^2 df}{1-r^2}$
Chi Square		$r = \sqrt{\chi^2/n}$	$\chi^2 = r^2 n$
Odds ratio	$d = \frac{ln(OR)}{1.81}$	$r = \frac{ln(OR)}{1.81\sqrt{(ln(OR)/1.81)^2+4}}$	$ln(OR) = \frac{3.62r}{\sqrt{1-r^2}}$
$r_{equivalent}$	r with probability p	$r = r_{equivalent}$	

The linear model is a regression model

1. $\hat{y} = \mu + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_1 * X_2 + ... + \epsilon$
2. Or more generally $\hat{y} = \mu + \boldsymbol{\beta}\mathbf{X} + \epsilon$ where $\boldsymbol{\beta}$ is a matrix of coefficients and $\mathbf{X}$ is a design matrix.
3. Analysis of variance is a special case where the **X** design matrix is a orthogonal set of weights.
4. Can use the `lm` or the `lmCor` functions to find the coefficients.
5. `lm` is in core-R and also gives convenient diagnostics
   - `lm` requires complete data and does not automatically zero-center interaction terms
   - `lmCor` will work with incomplete data, or the correlation matrix and by default zero centers before doing interaction products

Regression versus ANOVA

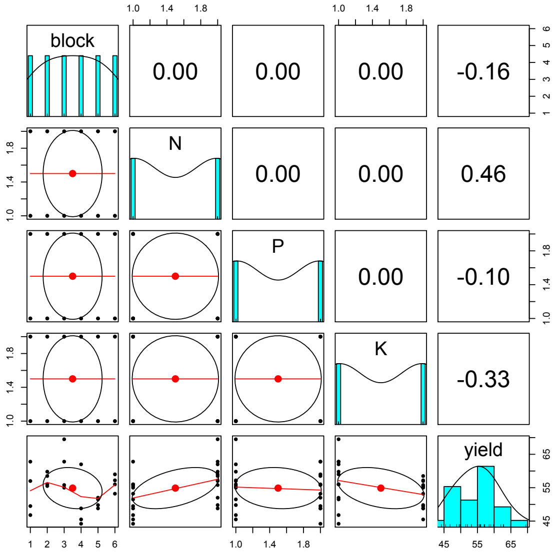
The npk data set is an example for anova. A classical N, P, K (nitrogen, phosphate, potassium) factorial experiment on the growth of peas conducted on 6 blocks. Each half of a fractional factorial design confounding the NPK interaction was used on 3 of the plots.

R code

```
describe(npk)  #raw data is categorical
NPK <- char2numeric(npk)  #convert to numeric
 describe(NPK)  #numeric
```

```
       vars  n  mean   sd median trimmed  mad  min  max range skew kurtosis   se
block*    1 24  3.50 1.74   3.50    3.50 2.22  1.0  6.0   5.0 0.00    -1.41 0.36
N*        2 24  1.50 0.51   1.50    1.50 0.74  1.0  2.0   1.0 0.00    -2.08 0.10
P*        3 24  1.50 0.51   1.50    1.50 0.74  1.0  2.0   1.0 0.00    -2.08 0.10
K*        4 24  1.50 0.51   1.50    1.50 0.74  1.0  2.0   1.0 0.00    -2.08 0.10
yield     5 24 54.88 6.17  55.65   54.75 6.15 44.2 69.5  25.3 0.24    -0.51 1.26

>
      vars  n  mean   sd median trimmed  mad  min  max range skew kurtosis   se
block    1 24  3.50 1.74   3.50    3.50 2.22  1.0  6.0   5.0 0.00    -1.41 0.36
N        2 24  1.50 0.51   1.50    1.50 0.74  1.0  2.0   1.0 0.00    -2.08 0.10
P        3 24  1.50 0.51   1.50    1.50 0.74  1.0  2.0   1.0 0.00    -2.08 0.10
K        4 24  1.50 0.51   1.50    1.50 0.74  1.0  2.0   1.0 0.00    -2.08 0.10
yield    5 24 54.88 6.17  55.65   54.75 6.15 44.2 69.5  25.3 0.24    -0.51 1.26
```

The npk data set



Compare `aov` to `lm`

1. aov works on factors and takes ‘products’ to show interactions
2. If the data are factors, `lm` will produce similar results
3. But, if the data are numeric, the results differ. Why?
4. First lets do the comparisons on factors

R code

```
summary(aov(yield ~ N * P * K, data=npk))
summary(lm(yield ~ N * P * K, data =npk))
```

```
> summary(aov(yield ~ N * P * K, data=npk))
            Df Sum Sq Mean Sq F value Pr(>F)
N            1  189.3  189.28   6.161 0.0245 *
P            1    8.4    8.40   0.273 0.6082
K            1   95.2   95.20   3.099 0.0975 .
N:P          1   21.3   21.28   0.693 0.4175
N:K          1   33.1   33.14   1.078 0.3145
P:K          1    0.5    0.48   0.016 0.9019
N:P:K        1   37.0   37.00   1.204 0.2887
Residuals   16  491.6   30.72
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(lm(yield ~ N * P * K, data =npk))
Call:
lm(formula = yield ~ N * P * K, data = npk)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  54.8750     1.1314  48.500   <2e-16 ***
N1            2.8083     1.1314   2.482   0.0245 *
P1           -0.5917     1.1314  -0.523   0.6082
K1           -1.9917     1.1314  -1.760   0.0975 .
N1:P1        -0.9417     1.1314  -0.832   0.4175
N1:K1        -1.1750     1.1314  -1.038   0.3145
P1:K1         0.1417     1.1314   0.125   0.9019
N1:P1:K1      1.2417     1.1314   1.097   0.2887
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 5.543 on 16 degrees of freedom
Multiple R-squared:  0.4391,        Adjusted R-squared:  0.1937
F-statistic: 1.789 on 7 and 16 DF,  p-value: 0.1586
```

But treating them numerically, the results differ

```
Call:
lm(formula = yield ~ N * P * K, data = npk)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  54.8750     1.1314  48.500   <2e-16 ***
N1            2.8083     1.1314   2.482   0.0245 *
P1           -0.5917     1.1314  -0.523   0.6082
K1           -1.9917     1.1314  -1.760   0.0975 .
N1:P1        -0.9417     1.1314  -0.832   0.4175
N1:K1        -1.1750     1.1314  -1.038   0.3145
P1:K1         0.1417     1.1314   0.125   0.9019
N1:P1:K1      1.2417     1.1314   1.097   0.2887
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 5.543 on 16 degrees of freedom
Multiple R-squared:  0.4391,        Adjusted R-squared:  0.1937
F-statistic: 1.789 on 7 and 16 DF,  p-value: 0.1586
> summary(lm(yield ~ N * P * K, data =NPK))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.900     35.779   0.081   0.9364
N             40.667     22.629   1.797   0.0912 .
P             25.967     22.629   1.148   0.2680
K             24.567     22.629   1.086   0.2937
N:P          -18.667     14.312  -1.304   0.2106
N:K          -19.600     14.312  -1.370   0.1898
P:K          -14.333     14.312  -1.002   0.3315
N:P:K          9.933      9.052   1.097   0.2887
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 5.543 on 16 degrees of freedom
Multiple R-squared:  0.4391,        Adjusted R-squared:  0.1937
F-statistic: 1.789 on 7 and 16 DF,  p-value: 0.1586
```

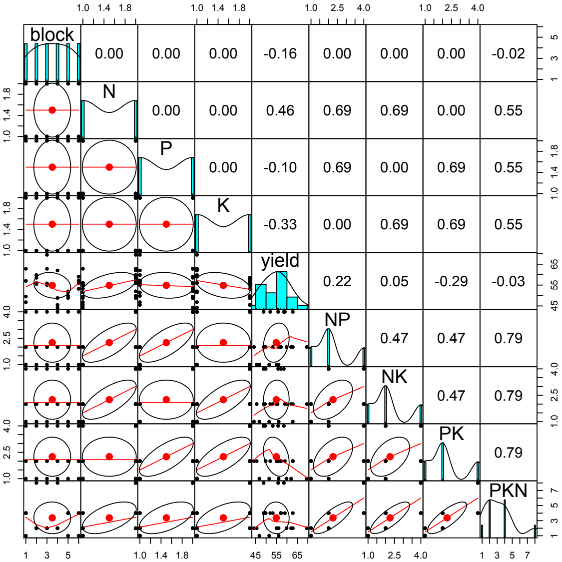
The problem with multiplication to produce interaction terms

1. An interaction is just the product of two variables (with the main effects removed)
2. But just taking the products will produce correlations between the main effects and the interactions.
3. We show this by finding the products and then their correlations

R code

```
 NP <- NPK$N * NPK$P
 NK <- NPK$N * NPK$K
 PK <- NPK$P * NPK$K
 PKN <- PK * NPK$N
 NPK.prods <- data.frame(NPK,NP,NK,PK,PKN)
pairs.panels(NPK.prods,gap=0)  #show the correlations,
      tighten up the figure
```

Interactions as products



Interactions of products of centered data

1. If we center the data (subtract the mean from each variable) aka deviation scores
2. Then the products are uncorrelated with the main effects
3. We can do this using the `scale` function
4. By default `scale` also standardizes (divides by the standard deviation).
5. To keep the data in the same metric as the raw data, we do not standardiize
6. Then do the regressions on the centered (with products) data

Center the data

R code

```
centered.NPK <- scale(NPK,scale=FALSE)
 centered.NPK <- data.frame(scale(NPK,scale=FALSE))
 c.NP <- centered.NPK$N * centered.NPK$P
 c.NK <- centered.NPK$N * centered.NPK$K
 c.PK <- centered.NPK$P * centered.NPK$K
 c.PKN <- c.PK * centered.NPK$N
 center.prod <- data.frame(centered.NPK,c.NP,c.NK,c.PK,c.PKN)
 describe(center.prod)
 pairs.panels(center.prod) #show the results grapically
```

```
describe(center.prod)
      vars  n mean   sd median trimmed  mad    min   max range skew kurtosis   se
block    1 24    0 1.74   0.00    0.00 2.22  -2.50  2.50  5.00 0.00    -1.41 0.36
N        2 24    0 0.51   0.00    0.00 0.74  -0.50  0.50  1.00 0.00    -2.08 0.10
P        3 24    0 0.51   0.00    0.00 0.74  -0.50  0.50  1.00 0.00    -2.08 0.10
K        4 24    0 0.51   0.00    0.00 0.74  -0.50  0.50  1.00 0.00    -2.08 0.10
yield    5 24    0 6.17   0.77   -0.13 6.15 -10.67 14.62 25.30 0.24    -0.51 1.26
c.NP     6 24    0 0.26   0.00    0.00 0.37  -0.25  0.25  0.50 0.00    -2.08 0.05
c.NK     7 24    0 0.26   0.00    0.00 0.37  -0.25  0.25  0.50 0.00    -2.08 0.05
c.PK     8 24    0 0.26   0.00    0.00 0.37  -0.25  0.25  0.50 0.00    -2.08 0.05
c.PKN    9 24    0 0.13   0.00    0.00 0.19  -0.12  0.12  0.25 0.00    -2.08 0.03
> lowerCor(center.prod)
      block N     P     K     yield c.NP  c.NK  c.PK  c.PKN
block  1.00
N      0.00  1.00
P      0.00  0.00  1.00
K      0.00  0.00  0.00  1.00
yield -0.16  0.46 -0.10 -0.33  1.00
c.NP   0.00  0.00  0.00  0.00  0.16  1.00
```

Describe and show correlations

```
describe(center.prod)
      vars  n mean   sd median trimmed  mad    min   max range skew kurtosis   se
block    1 24    0 1.74   0.00    0.00 2.22  -2.50  2.50  5.00 0.00    -1.41 0.36
N        2 24    0 0.51   0.00    0.00 0.74  -0.50  0.50  1.00 0.00    -2.08 0.10
P        3 24    0 0.51   0.00    0.00 0.74  -0.50  0.50  1.00 0.00    -2.08 0.10
K        4 24    0 0.51   0.00    0.00 0.74  -0.50  0.50  1.00 0.00    -2.08 0.10
yield    5 24    0 6.17   0.77   -0.13 6.15 -10.67 14.62 25.30 0.24    -0.51 1.26
c.NP     6 24    0 0.26   0.00    0.00 0.37  -0.25  0.25  0.50 0.00    -2.08 0.05
c.NK     7 24    0 0.26   0.00    0.00 0.37  -0.25  0.25  0.50 0.00    -2.08 0.05
c.PK     8 24    0 0.26   0.00    0.00 0.37  -0.25  0.25  0.50 0.00    -2.08 0.05
c.PKN    9 24    0 0.13   0.00    0.00 0.19  -0.12  0.12  0.25 0.00    -2.08 0.03
> lowerCor(center.prod)
      block N     P     K     yield c.NP  c.NK  c.PK  c.PKN
block  1.00
N      0.00  1.00
P      0.00  0.00  1.00
K      0.00  0.00  0.00  1.00
yield -0.16  0.46 -0.10 -0.33  1.00
c.NP   0.00  0.00  0.00  0.00 -0.16  1.00
c.NK   0.00  0.00  0.00  0.00 -0.19  0.00  1.00
c.PK   0.00  0.00  0.00  0.00  0.02  0.00  0.00  1.00
c.PKN -0.29  0.00  0.00  0.00  0.21  0.00  0.00  0.00  1.00
```

We do the linear model on the centered data

R code

```
summary(lm(yield ~ N*P*K,data=center.prod))
```

```
summary(lm(yield ~ N*P*K,data=center.prod))

Call:
lm(formula = yield ~ N * P * K, data = center.prod)

Residuals:
    Min      1Q  Median      3Q     Max
-10.133  -4.133   1.250   3.125   8.467

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.269e-15  1.131e+00   0.000   1.0000
N            5.617e+00  2.263e+00   2.482   0.0245 *
P           -1.183e+00  2.263e+00  -0.523   0.6082
K           -3.983e+00  2.263e+00  -1.760   0.0975 .
N:P         -3.767e+00  4.526e+00  -0.832   0.4175
N:K         -4.700e+00  4.526e+00  -1.038   0.3145
P:K          5.667e-01  4.526e+00   0.125   0.9019
N:P:K        9.933e+00  9.052e+00   1.097   0.2887
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 5.543 on 16 degrees of freedom
Multiple R-squared:  0.4391,     Adjusted R-squared:  0.1937
F-statistic: 1.789 on 7 and 16 DF,  p-value: 0.1586
```

This is now the same as the original aov

```
summary(aov(yield ~ N*P*K,data=npk))
            Df Sum Sq Mean Sq F value Pr(>F)
N            1  189.3  189.28   6.161 0.0245 *
P            1    8.4    8.40   0.273 0.6082
K            1   95.2   95.20   3.099 0.0975 .
N:P          1   21.3   21.28   0.693 0.4175
N:K          1   33.1   33.14   1.078 0.3145
P:K          1    0.5    0.48   0.016 0.9019
N:P:K        1   37.0   37.00   1.204 0.2887
Residuals   16  491.6   30.72
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(lm(yield ~ N*P*K,data=center.prod))
Call:
lm(formula = yield ~ N * P * K, data = center.prod)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.269e-15  1.131e+00   0.000   1.0000
N            5.617e+00  2.263e+00   2.482   0.0245 *
P           -1.183e+00  2.263e+00  -0.523   0.6082
K           -3.983e+00  2.263e+00  -1.760   0.0975 .
N:P         -3.767e+00  4.526e+00  -0.832   0.4175
N:K         -4.700e+00  4.526e+00  -1.038   0.3145
P:K          5.667e-01  4.526e+00   0.125   0.9019
N:P:K        9.933e+00  9.052e+00   1.097   0.2887
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.543 on 16 degrees of freedom
Multiple R-squared:  0.4391,	Adjusted R-squared:  0.1937
F-statistic: 1.789 on 7 and 16 DF,  p-value: 0.1586
```

Centering using the `scale` function

In the previous example, we hand centered the data. The `scale` function will do this. By default, it will also standardize. We avoid this by setting the *scale* parameter to FALSE.
Unfortunately, `scale` returns a *matrix* and we want a `data.frame`. This is irritating, but easily solved.
We use the Garcia data set.

R code

```
centered.Garcia <- data.frame(scale(Garcia, scale=FALSE))
describe(centered.Garcia)
```

```
> centered.Garcia <- data.frame(scale(Garcia, scale=FALSE))
> describe(centered.Garcia)
         vars   n mean   sd median trimmed  mad   min  max range  skew kurtosis   se
protest     1 129    0 0.82  -0.03    0.01 1.48 -1.03 0.97  2.00 -0.06    -1.52 0.07
sexism      2 129    0 0.78   0.00   -0.02 0.74 -2.25 1.88  4.13  0.12    -0.32 0.07
anger       3 129    0 1.66  -1.12   -0.29 0.00 -1.12 4.88  6.00  1.29     0.26 0.15
liking      4 129    0 1.05   0.19    0.09 0.99 -4.64 1.36  6.00 -1.15     2.48 0.09
respappr    5 129    0 1.35   0.38    0.12 1.11 -3.37 2.13  5.50 -0.75    -0.18 0.12
prot2       6 129    0 0.47   0.32    0.04 0.00 -0.68 0.32  1.00 -0.77    -1.41 0.04
```

A word of caution

1. aov and lm produce equivalent results *if* the design is balanced.
2. That is, if the IVs are represented proportionally. (no correlation between the $\mathbf{X_i}$)
3. Consider the case of the `Garcia` data set

```
lowerCor(Garcia)
         prtst sexsm anger likng rsppp prot2
protest   1.00
sexism   -0.02  1.00
anger    -0.31 -0.03  1.00
liking    0.17  0.09 -0.51  1.00
respappr  0.48  0.04 -0.53  0.49  1.00
prot2     0.86  0.04 -0.39  0.21  0.50  1.00
```

aov and lm not equivalent if design is unbalanced

R code

```
summary(aov(liking ~ prot2 +  sexism, data= Garcia))
summary(lm(liking ~ prot2 +  sexism, data= Garcia))
```

```
summary(aov(liking ~ prot2 +  sexism, data= Garcia))
             Df Sum Sq Mean Sq F value Pr(>F)
prot2         1   6.41   6.407   6.040 0.0153 *
sexism        1   0.97   0.969   0.913 0.3410
Residuals   126 133.66   1.061
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
> summary(lm(liking ~ prot2 +  sexism, data= Garcia))

Call:
lm(formula = liking ~ prot2 + sexism, data = Garcia)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3857 -0.6246  0.0599  0.7754  1.7954

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.7468     0.6110   7.768 2.41e-12 ***
prot2         0.4711     0.1949   2.417   0.0171 *
sexism        0.1111     0.1162   0.956   0.3410
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 1.03 on 126 degrees of freedom
Multiple R-squared:  0.0523,	Adjusted R-squared:  0.03726
F-statistic: 3.477 on 2 and 126 DF,  p-value: 0.03391
```

Even worse if look at interaction terms

R code

```
summary(aov(liking ~ prot2 *  sexism, data= Garcia))
summary(lm(liking ~ prot2 *  sexism, data= Garcia))
```

```
summary(aov(liking ~ prot2 *  sexism, data= Garcia))
               Df Sum Sq Mean Sq F value  Pr(>F)
prot2           1   6.41   6.407   6.553 0.01166 *
sexism          1   0.97   0.969   0.991 0.32139
prot2:sexism    1  11.45  11.451  11.713 0.00084 ***
Residuals     125 122.21   0.978
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
> summary(lm(liking ~ prot2 *  sexism, data= Garcia))

Call:
lm(formula = liking ~ prot2 * sexism, data = Garcia)
Residuals:
    Min      1Q  Median      3Q     Max
-3.9894 -0.6381  0.0478  0.7404  2.3650
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.7062     1.0449   7.375 1.99e-11 ***
prot2         -3.7727     1.2541  -3.008  0.00318 **
sexism        -0.4725     0.2038  -2.318  0.02205 *
prot2:sexism   0.8336     0.2436   3.422  0.00084 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 0.9888 on 125 degrees of freedom
Multiple R-squared:  0.1335,        Adjusted R-squared:  0.1127
```

Centering helps, but not if the DVs are correlated

R code

```
summary(aov(liking ~ prot2 *  sexism, data= Garcia))
summary(lm(liking ~ prot2 *  sexism, data= data.frame(scale(Garcia,sca
```

```
summary(aov(liking ~ prot2 *  sexism, data= Garcia))
              Df Sum Sq Mean Sq F value  Pr(>F)
prot2          1   6.41   6.407   6.553 0.01166 *
sexism         1   0.97   0.969   0.991 0.32139
prot2:sexism   1  11.45  11.451  11.713 0.00084 ***
Residuals    125 122.21   0.978
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
> summary(lm(liking ~ prot2 *  sexism, data= data.frame(scale(Garcia,scale=FALSE)) ))

Call:
lm(formula = liking ~ prot2 * sexism, data = data.frame(scale(Garcia,
    scale = FALSE)))

Residuals:
    Min      1Q  Median      3Q     Max
-3.9894 -0.6381  0.0478  0.7404  2.3650

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.01219    0.08713  -0.140  0.88899
prot2         0.49262    0.18722   2.631  0.00958 **
sexism        0.09613    0.11169   0.861  0.39102
prot2:sexism  0.83355    0.24356   3.422  0.00084 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

lmCor will do regressions and interactions as well

1. `lmCor` will work from raw data or correlation matrices
2. With raw data, it can find interactions
3. The syntax can be identical to `lm` or you can specify it by x and y
4. Compare

```
lmCor(yield ~ N * P * K, data = npk)
 with
 lm(yield ~ N * P * K, data = as.data.frame(scale(NPK)))
```

lmCor versus lm

```
lmCor(yield~ N*P*K,data= npk) #note, it will work on the factor level data as well
Call: lmCor(y = yield ~ N * P * K, data = npk)
Multiple Regression from raw data
 DV =  yield
            slope   se     t     p lower.ci upper.ci VIF
(Intercept)  0.00 0.19  0.00 1.000    -0.40     0.40   1
N            0.46 0.19  2.48 0.025     0.07     0.86   1
P           -0.10 0.19 -0.52 0.610    -0.49     0.30   1
K           -0.33 0.19 -1.76 0.097    -0.73     0.07   1
N*P         -0.16 0.19 -0.83 0.420    -0.55     0.24   1
N*K         -0.19 0.19 -1.04 0.310    -0.59     0.20   1
P*K          0.02 0.19  0.13 0.900    -0.37     0.42   1
N*P*K        0.21 0.19  1.10 0.290    -0.19     0.60   1
Residual Standard Error =  0.9  with  16  degrees of freedom
 Multiple Regression
        R   R2  Ruw R2uw Shrunken R2 SE of R2 overall F df1 df2     p
yield 0.66 0.44 0.56 0.31        0.19      0.1      1.79   7  16 0.159

lm(formula = yield ~ N * P * K, data = as.data.frame(scale(NPK)))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.195e-16  1.833e-01   0.000   1.0000
N            4.647e-01  1.872e-01   2.482   0.0245 *
P           -9.791e-02  1.872e-01  -0.523   0.6082
K           -3.296e-01  1.872e-01  -1.760   0.0975 .
N:P         -1.592e-01  1.913e-01  -0.832   0.4175
N:K         -1.986e-01  1.913e-01  -1.038   0.3145
P:K          2.395e-02  1.913e-01   0.125   0.9019
N:P:K        2.144e-01  1.954e-01   1.097   0.2887
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.898 on 16 degrees of freedom
```

Interactions are hard to visualize

1. Main effects (in anova terms) are just linear relationships
2. These may be shown by straight lines
3. Two main effects may be shown by two parallel lines
4. Interactions are non-parallel lines.
5. Lets use the Garcia et al. (2010) data to show this (in *psychTools* as `Garcia`).

Garcia data set

1. Garcia et al. (2010) report data for 129 subjects on the effects of perceived sexism on anger and liking of women's reactions to ingroup members who protest discrimination. This data set is also used as the 'protest' data set by Hayes (2013) It is a useful example of mediation and moderation in regression. It may also be used as an example of plotting interactions.
2. The reaction of women to women who protest discriminatory treatment was examined in an experiment reported by Garcia et al. (2010). 129 women were given a description of sex discrimination in the workplace (a male lawyer was promoted over a clearly more qualified female lawyer). Subjects then read that the target lawyer felt that the decision was unfair. Subjects were then randomly assigned to three conditions: Control (no protest), Individual Protest ("They are treating me unfairly") , or Collective Protest ("The firm is is treating women unfairly").
3. We use lmCor to find the regressions with the 0 centered product term and do the graphics at the same time

The Garcia data set

R code

```
dim(Garcia)
describe(Garcia)
lowerCor(Garcia)
```

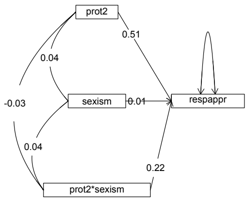
```
dim(Garcia)
[1] 129   6
> describe(Garcia)
         vars   n mean   sd median trimmed  mad  min max range  skew kurtosis   se
protest     1 129 1.03 0.82   1.00    1.04 1.48 0.00   2  2.00 -0.06    -1.52 0.07
sexism      2 129 5.12 0.78   5.12    5.10 0.74 2.87   7  4.13  0.12    -0.32 0.07
anger       3 129 2.12 1.66   1.00    1.84 0.00 1.00   7  6.00  1.29     0.26 0.15
liking      4 129 5.64 1.05   5.83    5.73 0.99 1.00   7  6.00 -1.15     2.48 0.09
respappr    5 129 4.87 1.35   5.25    4.98 1.11 1.50   7  5.50 -0.75    -0.18 0.12
prot2       6 129 0.68 0.47   1.00    0.72 0.00 0.00   1  1.00 -0.77    -1.41 0.04
> lowerCor(Garcia)
         prtst sexsm anger likng rsppp prot2
protest   1.00
sexism   -0.02  1.00
anger    -0.31 -0.03  1.00
liking    0.17  0.09 -0.51  1.00
respappr  0.48  0.04 -0.53  0.49  1.00
prot2     0.86  0.04 -0.39  0.21  0.50  1.00
```

Two analyses of Garcia–Center the data!

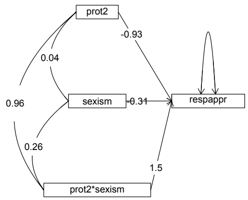
```
lmCor(respappr ~ prot2 * sexism ,data=Garcia ,main="Moderated regression (mean centered)")
Call: lmCor(y = respappr ~ prot2 * sexism, data = Garcia, main = "Moderated regression (mean c

Multiple Regression from raw data
 DV =  respappr
              slope   se    t       p lower.ci upper.ci VIF
(Intercept)    0.00 0.08 0.00 1.0e+00    -0.15     0.15   1
prot2          0.51 0.08 6.73 5.5e-10     0.36     0.65   1
sexism         0.01 0.08 0.18 8.6e-01    -0.14     0.16   1
prot2*sexism   0.22 0.08 2.87 4.8e-03     0.07     0.36   1
Residual Standard Error =  0.85  with  125  degrees of freedom
 Multiple Regression
           R  R2  Ruw R2uw Shrunken R2 SE of R2 overall F df1 df2        p
respappr 0.54 0.3 0.42 0.18        0.28     0.06     17.53   3 125 1.46e-09

>  lmCor(respappr ~ prot2 * sexism ,data=Garcia ,zero=FALSE,main="Moderated regression (not me
Call: lmCor(y = respappr ~ prot2 * sexism, data = Garcia, main = "Moderated regression (not me
    zero = FALSE)
Multiple Regression from raw data
 DV =  respappr
              slope   se     t      p lower.ci upper.ci   VIF
(Intercept)    0.00 0.08  0.00 1.0000    -0.15     0.15  1.00
prot2         -0.93 0.50 -1.85 0.0670    -1.93     0.06 44.99
sexism        -0.31 0.14 -2.24 0.0270    -0.58    -0.04  3.34
prot2*sexism   1.50 0.52  2.87 0.0048     0.47     2.53 48.14

Residual Standard Error =  0.85  with  125  degrees of freedom

 Multiple Regression
           R  R2  Ruw ../../images Shrunken R2 SE of R2 overall F df1 df2        p
respappr 0.54 0.3 0.45  0.2        0.28     0.06     17.53   3 125 1.46e-09
```

Comparing centered and non-centered interactions

Moderated regression (mean centered)



Moderated regression (not mean centered)



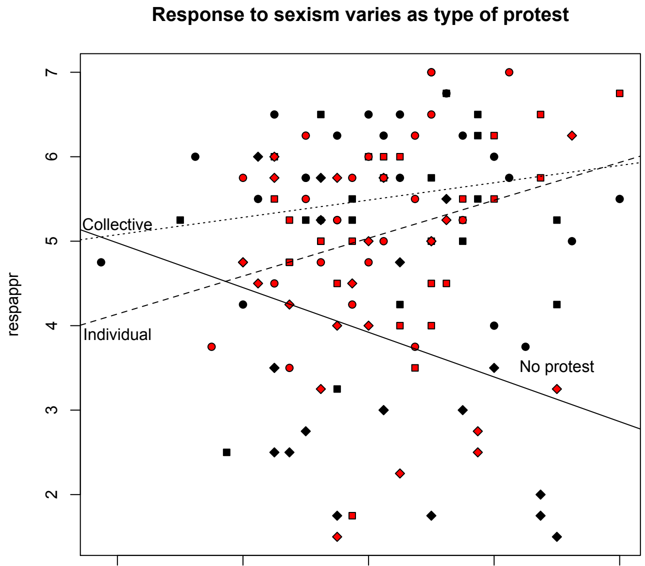
Plotting an interaction

1. Show the overall data as a function of group (different colors for different groups
2. Plot the regression lines separately for each group

R code

```
#demonstrate interaction plots
#first plot the data with a different color for each group
plot(respappr ~ sexism, pch = 23- protest,
      bg = c("black","red", "blue")[protest],
data=Garcia, main = "Response to sexism varies as type of protest")
#then, repeatedly draw a line for each regression slope
#use the abline function within the by function

by(Garcia,Garcia$protest, function(x) abline(lm(respappr ~ sexism,
   data =x),lty=c("solid","dashed","dotted")[x$protest+1]))
#Put in the labels for the graph
#the parameters are the x and y coordinates, followed by text to show
text(6.5,3.5,"No protest")
text(3,3.9,"Individual")
text(3,5.2,"Collective")
```

Showing an interaction



Can do the same interaction plot using `lmCor`

`lmCor` is meant to mimic `lm` for many of the results. The difference is in the default values. We adjust those to get the right result.

R code

```
#demonstrate interaction plots
#first plot the data with a different color for each group
plot(respappr ~ sexism, pch = 23- protest,
      bg = c("black","red", "blue")[protest],
data=Garcia, main = "Response to sexism varies as type of protest")
#then, repeatedly draw a line for each regression slope
#use the abline function within the by function
by(Garcia,Garcia$protest, function(x) abline(lmCor(respappr ~ sexism,
   data =x, plot=FALSE, std=FALSE)   #note that set these two paramete
         ,lty=c("solid","dashed","dotted")[x$protest+1]))
#Put in the labels for the graph
##the parameters are the x and y coordinates, followed by text to show

text(6.5,3.5,"No protest")
text(3,3.9,"Individual")
text(3,5.2,"Collective")
```

Athenstaedt, U. (2003). On the content and structure of the gender role self-concept: Including gender-stereotypical behaviors in addition to traits. *Psychology of Women Quarterly*, 27(4):309–318.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. L. Erlbaum Associates, Hillsdale, N.J., 2nd ed edition.

Del Giudice, M. (2009). On the real magnitude of psychological sex differences. *Evolutionary Psychology*, 7(2):147470490900700209.

Del Giudice, M., Booth, T., and Irwing, P. (2012). The distance between mars and venus: measuring global sex differences in personality. *PloS one*, 7(1):e29265–e29265.

Eagly, A. H. and Revelle, W. (2022). Understanding the Magnitude of Psychological Differences Between Women and Men Requires Seeing the Forest and the Trees. *Perspectives on Psychological Science*, 17(5):1339–1358.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263.

Galton, F. (1888). Co-relations and their measurement. *Proceedings of the Royal Society. London Series*, 45:135–145.

Garcia, D. M., Schmitt, M. T., Branscombe, N. R., and Ellemers, N. (2010). Women's reactions to ingroup members who protest discriminatory treatment: The importance of beliefs about inequality and response appropriateness. *European Journal of Social Psychology*, 40(5):733–745.

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford Press, New York.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Science (India)*, II(1):49–55.

McLachlan, G. J. (1999). Mahalanobis distance. *Resonance*, 4(6):20–26.

Pearson, K. (1896). Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. *Philisopical Transactions of the Royal Society of London. Series A*, 187:254–318.

Pearson, K. and Heron, D. (1913). On theories of association. *Biometrika*, 9(1/2):159–315.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Student (1908). The probable error of a mean. *Biometrika*, 6(1):1–25.