An introduction to Psychometric Theory
Correlation & Regression

William Revelle

Department of Psychology
Northwestern University
Evanston, Illinois USA

NORTHWESTERN UNIVERSITY

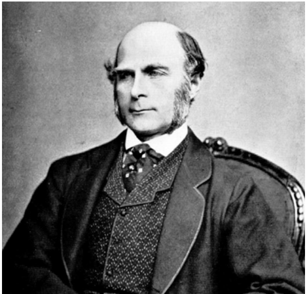
April, 2025

Outline

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.

https://personality-project.org/revelle/publications/galton.pdf(Revelle, 2015b)



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.

https://personality-project.org/revelle/publications/spearman.pdf (Revelle, 2015a)

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 values
```

parent \ child	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

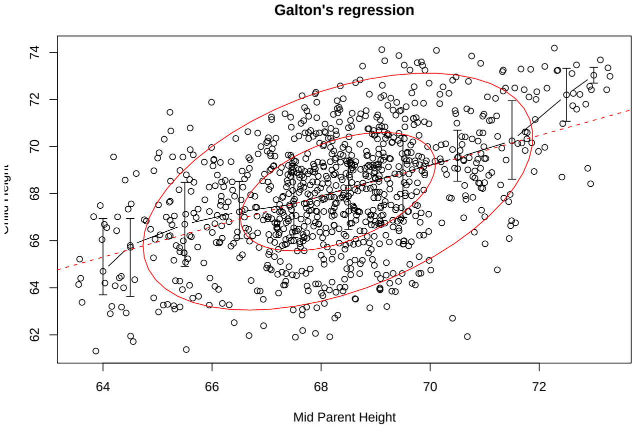
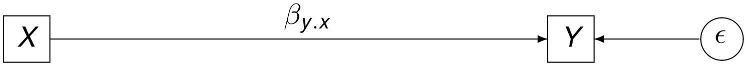


Figure: Galton's data can be plotted to show the relationships between mid parent and child heights. Because the original data are grouped, the data points have been *jittered* to emphasize the density of points along the median. The bars connect the first, 2nd (median) and third quartiles. The dashed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

Bivariate Regression

X Y ϵ



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

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

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

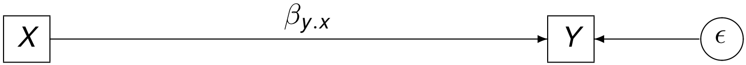
$$\sum(\epsilon^2) = \sum(y - \hat{y})^2 = \sum(y - \beta_{y.x}x)^2 = \sum(y^2 - 2y\beta_{y.x}x + (\beta_{y.x}x)^2)$$

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

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

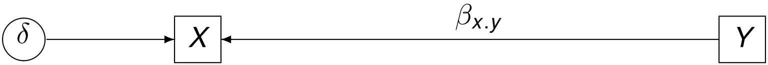
Bivariate Regression

δ X Y ϵ



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

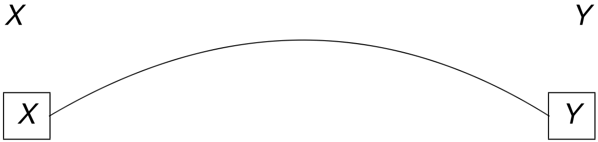
$$\beta_{y.x} = \frac{\sigma_{xy}}{\sigma_x^2}$$



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

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

Bivariate Correlation is the geometric average of the two regressions



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

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

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

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

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

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

The variance and the variance of a composite

1. If $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ are vectors of N observations centered around their mean (that is, deviation scores) their variances are $V_{x_1} = \sum x_{1_i}^2/(N-1)$ and $V_{x_{2_i}} = \sum x_{2_i}^2/(N-1)$, or, in matrix terms $V_{\boldsymbol{x}_1} = \boldsymbol{x_1'x_1}/(N-1)$ and $V_{\boldsymbol{x}2} = \boldsymbol{x_2'x_2}/(N-1)$.
2. The variance of the composite made up of the sum of the corresponding scores, $\boldsymbol{x_1} + \boldsymbol{x_2}$ is just

$$V_{(\boldsymbol{x_1}+\boldsymbol{x_2})} = \frac{\sum(x_1+x_2)^2}{N-1} = \frac{\sum x_{1_i}^2 + \sum x_{2_i}^2 + 2\sum x_{1_i}x_{2_i}}{N-1} = \frac{(\boldsymbol{x_1}+\boldsymbol{x_2})'(\boldsymbol{x_1}+\boldsymbol{x_2})}{N-1}. \quad (1)$$

Or, more generally,

$$\boldsymbol{S} = \begin{pmatrix} V_{x1} & C_{x1x2} & \cdots & C_{x1xn} \\ C_{x1x2} & V_{x2} & & C_{x2xn} \\ \vdots & & \ddots & \vdots \\ C_{x1xn} & C_{x2xn} & \cdots & V_{xn} \end{pmatrix}$$

Sums as matrix products

$$V_{\boldsymbol{X}} = \sum \frac{\boldsymbol{X}'\boldsymbol{X}}{N-1} = \frac{\mathbf{1}'(\boldsymbol{X}'\boldsymbol{X})\mathbf{1}}{N-1}.$$

$$V_{\boldsymbol{Y}} = \sum \frac{\boldsymbol{Y}'\boldsymbol{Y}}{N-1} = \frac{\mathbf{1}'(\boldsymbol{Y}'\boldsymbol{Y})\mathbf{1}}{N-1}$$

and

$$C_{\boldsymbol{XY}} = \sum \frac{\boldsymbol{X}'\boldsymbol{Y}}{N-1} = \frac{\mathbf{1}'(\boldsymbol{X}'\boldsymbol{Y})\mathbf{1}}{N-1}$$

Use R

Get the data from a remote data source

A nice feature of R is that you can read from remote data sets. The example dataset is on the personality-project.org server. Get it and describe it.

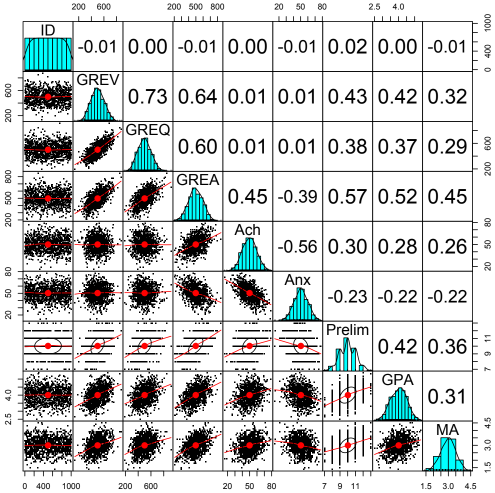
```
>  datafilename="https://personality-project.org/r/datasets/psychometrics.prob2.txt"
> mydata =read.file(datafilename)   #read the data file
> describe(mydata,skew=FALSE)
```

	var	n	mean	sd	median	trimmed	mad	min	max	range	se
ID	1	1000	500.50	288.82	500.50	500.50	370.65	1.0	1000.00	999.00	9.13
GREV	2	1000	499.77	106.11	497.50	498.75	106.01	138.0	873.00	735.00	3.36
GREQ	3	1000	500.53	103.85	498.00	498.51	105.26	191.0	914.00	723.00	3.28
GREA	4	1000	498.13	100.45	495.00	498.67	99.33	207.0	848.00	641.00	3.18
Ach	5	1000	49.93	9.84	50.00	49.88	10.38	16.0	79.00	63.00	0.31
Anx	6	1000	50.32	9.91	50.00	50.43	10.38	14.0	78.00	64.00	0.31
Prelim	7	1000	10.03	1.06	10.00	10.02	1.48	7.0	13.00	6.00	0.03
GPA	8	1000	4.00	0.50	4.02	4.01	0.53	2.5	5.38	2.88	0.02
MA	9	1000	3.00	0.49	3.00	3.00	0.44	1.4	4.50	3.10	0.02

Plot it using the `pairs.panels` function.

Use the pairs.panels function to show a splom plot (use gap=0 and pch='.').

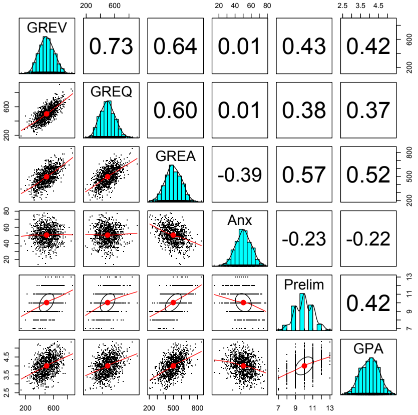
>pairs.panels(mydata,pch=".",gap=0) #pch='.' makes for a cleaner plot



Plot a subset of the data using the c() function (concatenate).

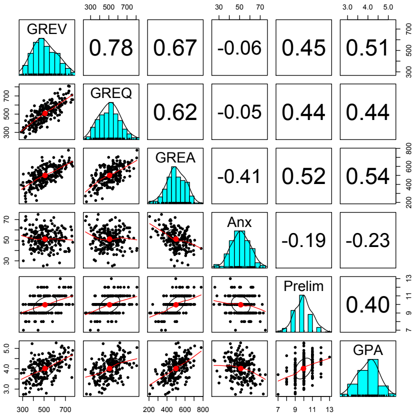
Use the pairs.panels function to show a splom plot. Select a subset of variables using the c() function.

```
>pairs.panels(mydata[c(2:4,6:8)],pch='.')
```



Do this for the first 200 subjects

```
> pairs.panels(mydata[mydata$ID < 200,c(2:4,6:8)])
```



0 center the data

In order to interpret interaction terms along with main effects in regressions, it is necessary to 0 center the data. We need to turn the result into a data.frame in order to use it in the regression(`lm`)function.

```
> cent <- data.frame(scale(mydata,scale=FALSE))
> describe(cent,skew=FALSE)
```

	var	n	mean	sd	median	trimmed	mad	min	max	range	se
ID	1	1000	0	288.82	0.00	0.00	370.65	-499.50	499.50	999.00	9.13
GREV	2	1000	0	106.11	-2.27	-1.02	106.01	-361.77	373.23	735.00	3.36
GREQ	3	1000	0	103.85	-2.53	-2.02	105.26	-309.53	413.47	723.00	3.28
GREA	4	1000	0	100.45	-3.13	0.54	99.33	-291.13	349.87	641.00	3.18
Ach	5	1000	0	9.84	0.07	-0.05	10.38	-33.93	29.07	63.00	0.31
Anx	6	1000	0	9.91	-0.32	0.11	10.38	-36.32	27.68	64.00	0.31
Prelim	7	1000	0	1.06	-0.03	0.00	1.48	-3.03	2.97	6.00	0.03
GPA	8	1000	0	0.50	0.02	0.00	0.53	-1.50	1.38	2.88	0.02
MA	9	1000	0	0.49	0.00	0.00	0.44	-1.60	1.50	3.10	0.02

The standard deviations and ranges have not changed. However, the means are all 0. We use the `scale` function with the scale=FALSE option.

The standardized data

Alternatively, we could standardize it.

```
> z.data <- data.frame(scale(my.data))
>  describe(z.data)
```

	var	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
ID	1	1000	0	1	0.00	0.00	1.28	-1.73	1.73	3.46	0.00	-1.20	0.03
GREV	2	1000	0	1	-0.02	-0.01	1.00	-3.41	3.52	6.93	0.09	-0.07	0.03
GREQ	3	1000	0	1	-0.02	-0.02	1.01	-2.98	3.98	6.96	0.22	0.08	0.03
GREA	4	1000	0	1	-0.03	0.01	0.99	-2.90	3.48	6.38	-0.02	-0.06	0.03
Ach	5	1000	0	1	0.01	-0.01	1.05	-3.45	2.95	6.40	0.00	0.02	0.03
Anx	6	1000	0	1	-0.03	0.01	1.05	-3.67	2.79	6.46	-0.14	0.14	0.03
Prelim	7	1000	0	1	-0.02	0.00	1.40	-2.86	2.81	5.67	-0.02	-0.01	0.03
GPA	8	1000	0	1	0.03	0.01	1.06	-3.00	2.74	5.74	-0.07	-0.29	0.03
MA	9	1000	0	1	0.01	0.01	0.90	-3.23	3.04	6.27	-0.07	-0.09	0.03

Or, we can standardize it by dividing though by the standard deviation. We use the `scale` function to do this for us.

Show how the correlations do not change with standardization

Find the correlations using the `lowerCor` function. This, by default, uses pairwise Pearson correlations and rounds to two decimals. Compare with the standard `cor` function.

```
> lowerCor(my.data)

       ID    GREV  GREQ  GREA  Ach   Anx
Prelm GPA   MA
ID      1.00
GREV   -0.01  1.00
GREQ    0.00  0.73  1.00
GREA   -0.01  0.64  0.60  1.00
Ach     0.00  0.01  0.01  0.45  1.00
Anx    -0.01  0.01  0.01 -0.39 -0.56  1.00
Prelim  0.02  0.43  0.38  0.57  0.30 -0.23
1.00
GPA     0.00  0.42  0.37  0.52  0.28 -0.22
0.42  1.00
MA     -0.01  0.32  0.29  0.45  0.26 -0.22
0.36  0.31  1.00
```

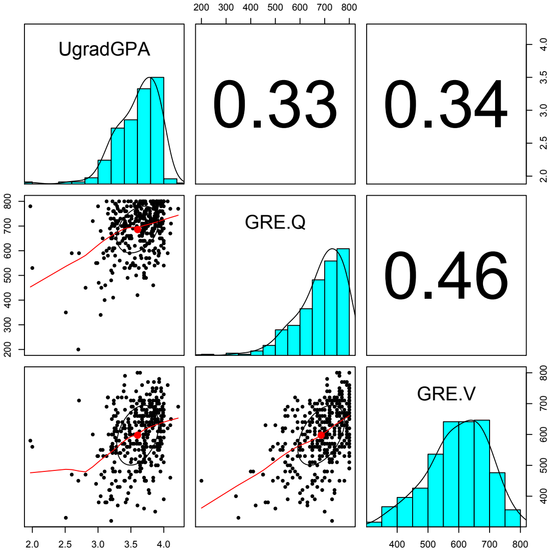
```
> lowerCor(z.data)

       ID    GREV  GREQ  GREA  Ach   Anx
Prelm GPA   MA
ID      1.00
GREV   -0.01  1.00
GREQ    0.00  0.73  1.00
GREA   -0.01  0.64  0.60  1.00
Ach     0.00  0.01  0.01  0.45  1.00
Anx    -0.01  0.01  0.01 -0.39 -0.56  1.00
Prelim  0.02  0.43  0.38  0.57  0.30 -0.23
1.00
GPA     0.00  0.42  0.37  0.52  0.28 -0.22
0.42  1.00
MA     -0.01  0.32  0.29  0.45  0.26 -0.22
0.36  0.31  1.00
```

Show that the two matrices do not differ using the `lowerUpper` function

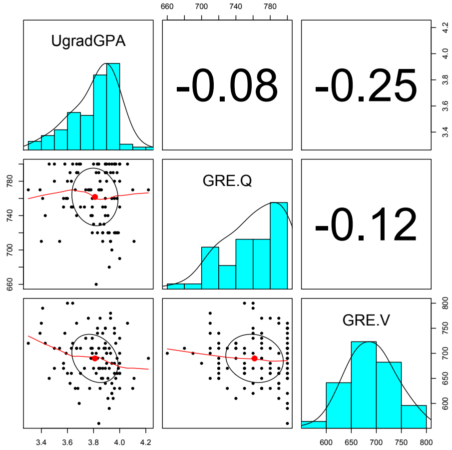
```
r <- lowerCor(my.data)    #find the original correlations
z <- lowerCor(z.data)     #find the z transformed correlations
lu <- lowerUpper(r,z,diff=TRUE)  #combine into one matrix and take the difference

round(lu,2)
```

	ID	GREV	GREQ	GREA	Ach	Anx	Prelim	GPA	MA
ID	NA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
GREV	−0.01	NA	0.00	0.00	0.00	0.00	0.00	0.00	0
GREQ	0.00	0.73	NA	0.00	0.00	0.00	0.00	0.00	0
GREA	−0.01	0.64	0.60	NA	0.00	0.00	0.00	0.00	0
Ach	0.00	0.01	0.01	0.45	NA	0.00	0.00	0.00	0
Anx	−0.01	0.01	0.01	−0.39	−0.56	NA	0.00	0.00	0
Prelim	0.02	0.43	0.38	0.57	0.30	−0.23	NA	0.00	0
GPA	0.00	0.42	0.37	0.52	0.28	−0.22	0.42	NA	0
MA	−0.01	0.32	0.29	0.45	0.26	−0.22	0.36	0.31	NA

Scatter Plot Matrix showing correlation and LOESS regression

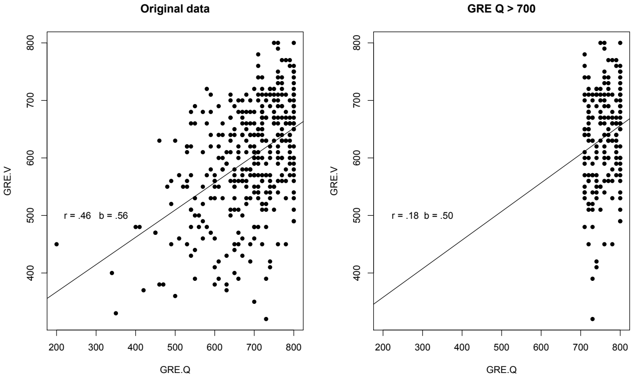


The effect of selection on the correlation



- 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

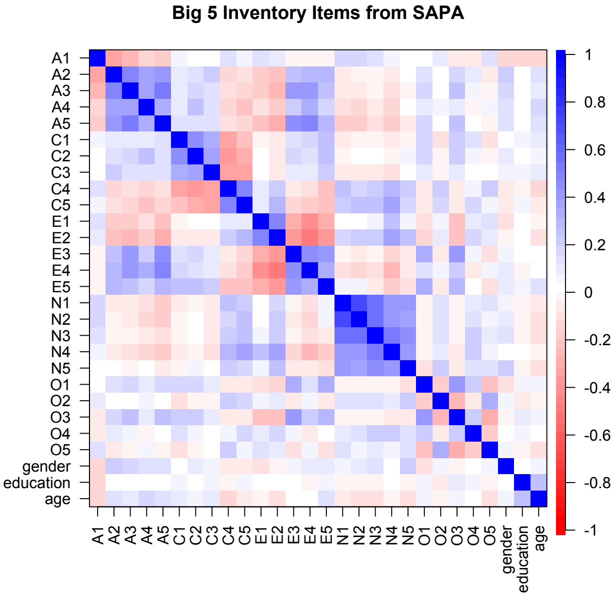


Although the correlation is very sensitive, regression slopes are relatively insensitive to restriction of range.

R code for regression figures

```
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	Assumptions
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	bivariate normality
Polychoric	r_{pc}	categorical	categorical	bivariate normality

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

$$n\sum X_i Y_i - \sum X_i \sum Y_i = nA - R_1 C_1$$

$$\phi = \frac{AD - BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}. \quad (2)$$

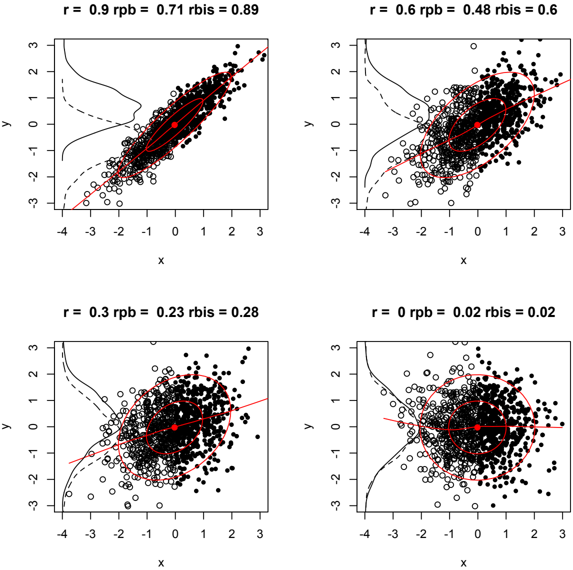
Correlation size $\neq$ causal importance

Table: The relationship between sex and pregnancy (hypothetical data)

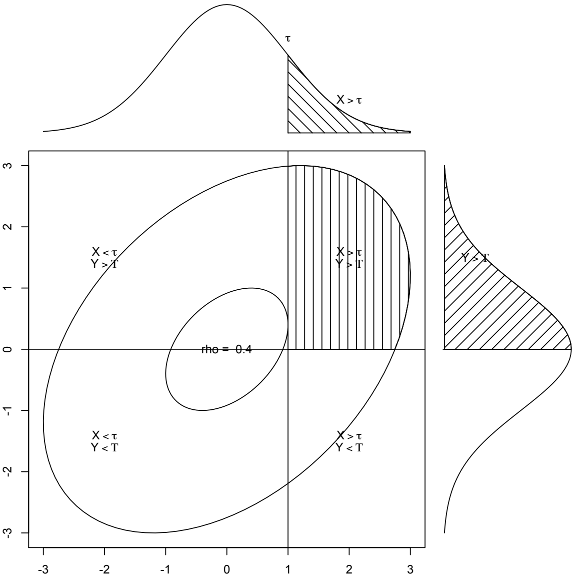
	Pregnant	Not Pregnant	Total
Intercourse	2	1,041	1,043
No intercourse	0	6,257	6,257
Total	2	7,298	7,300
Phi	.04		

```
> sex <- c(2, 1041,0,6257)
> phi(sex)

[1] 0.04
```

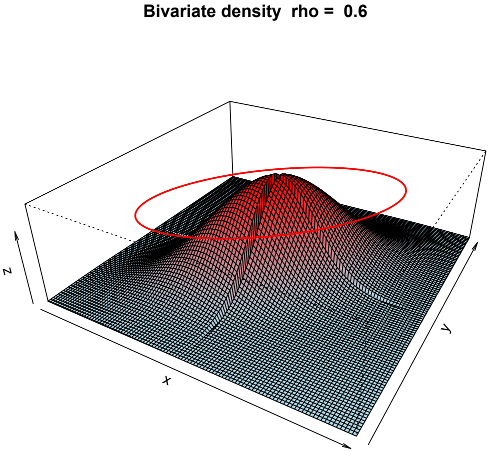
The biserial correlation estimates the latent correlation



The tetrachoric correlation estimates the latent correlation



The tetrachoric correlation estimates the latent correlation



Correlation size $\neq$ causal importance – tetrachoric correlation

Table: The relationship between sex and pregnancy (hypothetical data)

	Pregnant	Not Pregnant	Total
Intercourse	2	1,041	1,043
No intercourse	0	6,257	6,257
Total	2	7,298	7,300
Phi	.04	ρ_{tet}	.95

```
> sex <- c(2, 1041,0,6257)
> phi(sex)

[1] 0.04

> tetrachoric(sex,correct=FALSE)

Call: tetrachoric(x = sex, correct = FALSE)
tetrachoric correlation
[1] 0.95

 with tau of
[1] -3.5 -1.1
```

Pearson r versus tetrachoric correlation on dichotomous ability data

```
> tet <- tetrachoric(ability)
Loading required package: mvtnorm
Loading required package: parallel
> per <- lowerCor(ability)
> per.tet <- lowerUpper(tet$rho,per)
> per.tet.diff <- lowerUpper(tet$rho,per,diff=TRUE)
> round(per.tet[1:8,1:8],2)
```

```
          reason.4 reason.16 reason.17 reason.19 letter.7 letter.33 letter.34 letter.58
reason.4        NA      0.28      0.40      0.30     0.28      0.23      0.29      0.29
reason.16     0.45        NA      0.32      0.25     0.27      0.20      0.26      0.21
reason.17     0.61      0.51        NA      0.34     0.29      0.26      0.29      0.29
reason.19     0.46      0.40      0.53        NA     0.25      0.25      0.27      0.25
letter.7      0.45      0.43      0.47      0.40       NA      0.34      0.40      0.33
letter.33     0.37      0.32      0.42      0.39     0.52        NA      0.37      0.28
letter.34     0.46      0.41      0.47      0.43     0.60      0.56        NA      0.32
letter.58     0.47      0.35      0.48      0.40     0.51      0.43      0.50        NA
> round(per.tet.diff[1:8,1:8],2)
          reason.4 reason.16 reason.17 reason.19 letter.7 letter.33 letter.34 letter.58
reason.4        NA      0.17      0.21      0.17     0.16      0.14      0.17      0.18
reason.16     0.45        NA      0.19      0.15     0.16      0.13      0.16      0.14
reason.17     0.61      0.51        NA      0.19     0.18      0.16      0.18      0.19
reason.19     0.46      0.40      0.53        NA     0.14      0.14      0.15      0.15
letter.7      0.45      0.43      0.47      0.40       NA      0.18      0.20      0.18
letter.33     0.37      0.32      0.42      0.39     0.52        NA      0.19      0.15
letter.34     0.46      0.41      0.47      0.43     0.60      0.56        NA      0.18
letter.58     0.47      0.35      0.48      0.40     0.51      0.43      0.50        NA
```

Pearson r versus polychoric correlation on 6 alternative BFI data

```
> poly <- polychoric(bfi[1:10])
> pearson <- cor(bfi[1:10],use="pairwise")
> poly.pear <- lowerUpper(poly$rho,pearson)
> poly.pear.diff <- lowerUpper(poly$rho,pearson,diff=TRUE)
> poly.pear
```

```
> round(poly.pear,2)
      A1    A2    A3    A4    A5    C1    C2    C3    C4    C5
A1    NA -0.34 -0.27 -0.15 -0.18  0.03  0.02 -0.02  0.13  0.05
A2 -0.41    NA  0.49  0.34  0.39  0.09  0.14  0.19 -0.15 -0.12
A3 -0.32  0.56    NA  0.36  0.50  0.10  0.14  0.13 -0.12 -0.16
A4 -0.18  0.39  0.41    NA  0.31  0.09  0.23  0.13 -0.15 -0.24
A5 -0.23  0.45  0.57  0.36    NA  0.12  0.11  0.13 -0.13 -0.17
C1  0.00  0.12  0.12  0.11  0.16    NA  0.43  0.31 -0.34 -0.25
C2  0.01  0.16  0.16  0.27  0.14  0.48    NA  0.36 -0.38 -0.30
C3 -0.02  0.23  0.16  0.17  0.15  0.34  0.40    NA -0.34 -0.34
C4  0.15 -0.19 -0.16 -0.20 -0.17 -0.40 -0.43 -0.38    NA  0.48
C5  0.06 -0.16 -0.19 -0.28 -0.20 -0.29 -0.33 -0.38  0.53    NA
> round(poly.pear.diff,2)
      A1    A2    A3    A4    A5    C1    C2    C3    C4    C5
A1    NA -0.07 -0.06 -0.03 -0.05 -0.02 -0.01  0.00  0.02  0.01
A2 -0.41    NA  0.07  0.05  0.06  0.02  0.02  0.03 -0.05 -0.03
A3 -0.32  0.56    NA  0.05  0.07  0.03  0.02  0.03 -0.04 -0.03
A4 -0.18  0.39  0.41    NA  0.05  0.02  0.04  0.04 -0.04 -0.04
A5 -0.23  0.45  0.57  0.36    NA  0.04  0.03  0.02 -0.04 -0.03
C1  0.00  0.12  0.12  0.11  0.16    NA  0.06  0.04 -0.06 -0.04
C2  0.01  0.16  0.16  0.27  0.14  0.48    NA  0.04 -0.05 -0.03
C3 -0.02  0.23  0.16  0.17  0.15  0.34  0.40    NA -0.04 -0.04
C4  0.15 -0.19 -0.16 -0.20 -0.17 -0.40 -0.43 -0.38    NA  0.05
C5  0.06 -0.16 -0.19 -0.28 -0.20 -0.29 -0.33 -0.38  0.53    NA
```

Spearman vs. Pearson on BFI data

The lower off diagonal are the Spearman correlations, the upper off diagonal report the differences between Spearman and Pearson correlations. This

```
> spear <- cor(bfi[1:10],use="pairwise",method="spearman")
> spear.pear <- lowerUpper(spear,pearson,diff=TRUE)
> round(spear.pear,2)
```

	A1	A2	A3	A4	A5	C1	C2	C3	C4	C5
A1	NA	−0.03	−0.03	−0.01	−0.04	−0.05	−0.03	−0.02	0.02	0.01
A2	−0.37	NA	0.02	0.00	0.01	0.02	0.01	0.01	−0.03	−0.03
A3	−0.30	0.50	NA	0.00	0.03	0.02	0.01	0.02	−0.03	−0.02
A4	−0.16	0.34	0.36	NA	0.01	0.01	0.02	0.02	−0.03	−0.01
A5	−0.22	0.40	0.53	0.31	NA	0.02	0.02	0.01	−0.03	−0.02
C1	−0.02	0.11	0.12	0.10	0.15	NA	0.02	0.01	−0.04	−0.01
C2	−0.01	0.14	0.15	0.25	0.13	0.45	NA	0.01	−0.02	0.00
C3	−0.04	0.21	0.16	0.15	0.14	0.32	0.37	NA	−0.01	−0.01
C4	0.15	−0.18	−0.16	−0.18	−0.16	−0.38	−0.40	−0.35	NA	0.01
C5	0.06	−0.15	−0.18	−0.26	−0.19	−0.26	−0.30	−0.35	0.49	NA

Comments on these alternative correlations

1. The assumption is that there was an underlying bivariate, normal distribution that was somehow artificially dichotomized.
2. But some things are in fact dichotomous, not normally distributed
   - Alive/Dead
   - Vacinated/Not vacinated
3. polychoric and tetrachoric correlations are found by iteratively fitting bivariate normal distributions with varying correlations until the best fit for a n x n table is found.
4. This is done using the `tetrachoric` or `polychoric` functions. They are not fast! (In comparison to Pearson r), but have been pretty well optimized.

Pearson on correlation

> *It is this conception of correlation be- tween two occur- rences embracing all re- lationshisp from absolute inde- pendence to complete dependence, which is the wider category by which we have to re- place the old idea of causation. (Pearson, 1910, p 157)*

But

> *Such high correlations as arise from com- mon growth or decline with time, when interpreted as causal or sem- icausal rela- tionships, are in our opinion perfectly idle, indeed are only too apt to be mis- chievous, and we shall reach nothing, or less than nothing-knighthoods-by the in- vestigation of them.*

Cautions about correlations–The Anscombe data set

Consider the following 8 variables

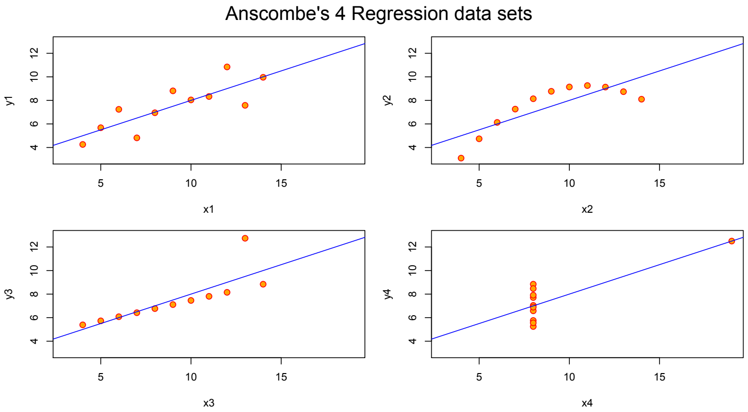
	var	n	**mean**	**sd**	**median**	trimmed	**mad**	**min**	**max**	**range**	skew	kurtosis	**se**
x1	1	11	9.0	3.32	9.00	9.00	4.45	4.00	14.00	10.00	0.00	–1.20	1.
x2	2	11	9.0	3.32	9.00	9.00	4.45	4.00	14.00	10.00	0.00	–1.20	1.
x3	3	11	9.0	3.32	9.00	9.00	4.45	4.00	14.00	10.00	0.00	–1.20	1.
x4	4	11	9.0	3.32	8.00	8.00	0.00	8.00	19.00	11.00	2.47	11.00	1.
y1	5	11	7.5	2.03	7.58	7.49	1.82	4.26	10.84	6.58	–0.05	–0.53	0.
y2	6	11	7.5	2.03	8.14	7.79	1.47	3.10	9.26	6.16	–0.98	0.85	0.
y3	7	11	7.5	2.03	7.11	7.15	1.53	5.39	12.74	7.35	1.38	4.38	0.
y4	8	11	7.5	2.03	7.04	7.20	1.90	5.25	12.50	7.25	1.12	3.15	0.

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
```

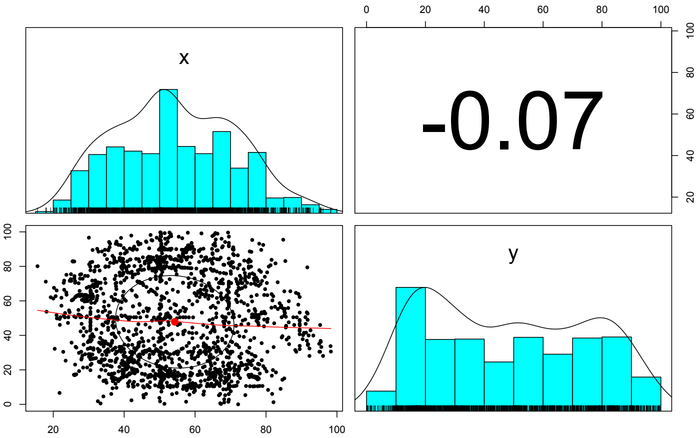
Cautions about correlations: Anscombe data set



A rather boring data set.

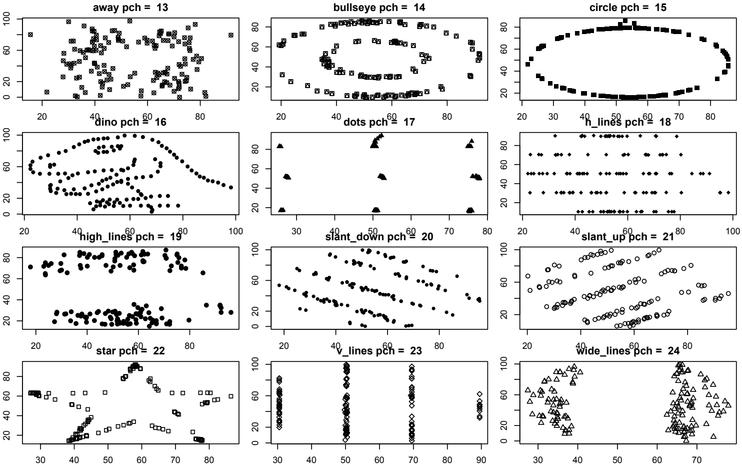
The overall plot of all the data shows no relationship

From Davies R, Locke S, D'Agostino McGowan L (2022). datasauRus: Datasets from the Datasaurus Dozen. R package version 0.1.6, <https://CRAN.R-project.org/package=datasauRus>.



	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
dataset*	1	1846	7.00	3.74	7.00	7.00	4.45	1.00	13.00	12.00	0.00	-1.22	0.09
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

Plotting data is very helpful

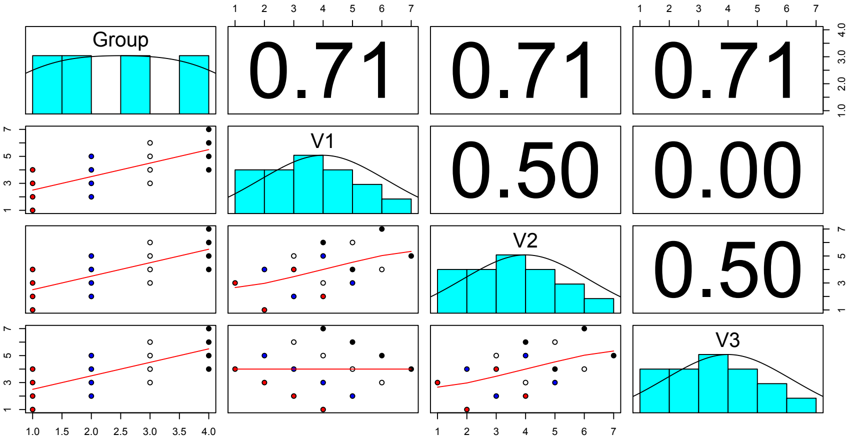


All sets $\mu_x = 54.26$, $\mu_y = 47.84$, $\sigma_x = 16.77 \sigma_y = 26.93$, $r = .07$

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



Bias, or just the Yule-Simpson Paradox?

Table: Hypothetical Admissions data showing sex discrimination

	Admit	Reject	Total
Male	40	10	50
Female	10	40	50
Total	50	50	100

Phi =(VP - HR*SR) /sqrt(HR*(1-HR)*(SR)*(1-SR)= .60 polychoric rho = .81

Aldrich (1995); Bickel et al. (1975); Simpson (1951); Yule (1903)

Covid mortality by gender in Belgium

For all ages, men have a higher mortality than women, but the overall mortality for women is higher.

Age groups	Men (%)	Women %)
0–24	0.00	0.0
25–44	0.02	0.01
45–64	0.29	0.14
65–74	2.92	1.61
75–84	5.56	3.35
85 and older	13.20	11.07
All ages	1.18	1.31

Wang and Rousseau (2021) appliy the Yule-Simpson problem to the question of how to aggregate citation statistics across journals or across fields or countries.

Calculate the ϕ and tetrachoric correlations

```
> admit <- c(40,10,10,40)
> phi(admit)

[1] 0.6

> phi2poly(.6,.5,.5)

[1] 0.8090178

> tetrachoric(admit)

Call: tetrachoric(x = admit)
tetrachoric correlation
[1] 0.81

 with tau of
[1] 0 0
```

1. Input the four cell counts
2. Find the ϕ coefficient
3. Convert this to a tetrachoric correlation by specifying the marginals
4. Or, just call tetrachoric with these cell entries

Sex discrimination by department shows opposite effect

Table: Hypothetical Admissions data showing sex discrimination

	Admit	Reject	Total
Male	40	10	50
Female	10	40	50
Total	50	50	100

Table: Males: unselective

	Admit	Reject	Total
Male	40	5	45
Female	5	0	5
Total	45	5	50
ϕ	-.11	ρ	-.95

Table: Females: selective

	Admit	Reject	Total
Male	0	5	5
Female	5	40	45
Total	5	45	50
ϕ	-.11	ρ	-.95

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	$b_{y.x} = \frac{Cxy}{\sigma_x^2}$	$r = b_{y.x} \frac{\sigma_x}{\sigma_y}$	$b_{y.x} = r \frac{\sigma_y}{\sigma_x}$
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	$r = r_{equivalent}$	

Correlation as the average of regressions

Galton's insight was that if both x and y were on the same scale with equal variability, then the slope of the line was the same for both predictors and was measure of the strength of their relationship. Galton (1886) converted all deviations to the same metric by dividing through by half the interquartile range, and Pearson (1896) modified this by converting the numbers to standard scores (i.e., dividing the deviations by the standard deviation). Alternatively, the geometric mean of the two slopes ($b_x y$ and $b_y x$) leads to the same outcome:

$$r_{xy} = \sqrt{b_{xy} b_{yx}} = \sqrt{\frac{(Cov_{xy} Cov_{yx}}{\sigma_x^2 \sigma_y^2}} = \frac{Cov_{xy}}{\sqrt{\sigma_x^2 \sigma_y^2}} = \frac{Cov_{xy}}{\sigma_x \sigma_y} \quad (3)$$

which is the same as the covariance of the standardized scores of X and Y.

$$r_{xy} = Cov_{z_x z_y} = Cov_{\frac{x}{\sigma_x} \frac{y}{\sigma_y}} = \frac{Cov_{xy}}{\sigma_x \sigma_y} \quad (4)$$

The slope $b_{y.x}$ was found so that it minimizes the sum of the squared residual, but what is it? That is, how big is the variance of the residual?

$$V_r = \sum_{i=1}^{n}(y - \hat{y})^2/n = \sum_{i=1}^{n}(y - b_{y.x}x)^2/n$$

$$V_r = \sum_{i=1}^{n}(y^2 + b_{y.x}^2 x^2 - 2b_{y.x}xy)/n$$

$$V_r = V_y + \frac{Cov_{xy}^2}{V_x} - 2\frac{Cov_{xy}^2}{V_x} = V_y - \frac{Cov_{xy}^2}{V_x}$$

$$V_r = V_y - r_{xy}^2 V_y = V_y(1 - r_{xy}^2) \tag{5}$$

That is, the *variance of the residual* in Y or the variance of the error of prediction of Y is the product of the original variance of Y and one minus the squared correlation between X and Y. The squared correlation between x and y is thus an index of the amount of variance in Y that is linearly predicted by X. This squared correlation is known as the *index of determination*.

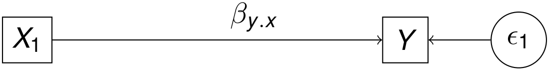
Variance and correlations

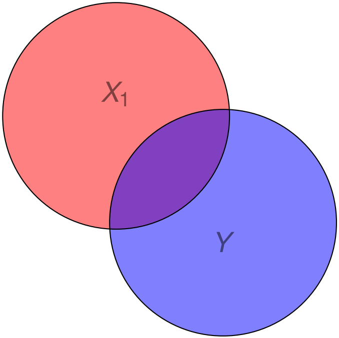
The various relationships between correlations, predicted scores, the variance of the predicted scores, and the variances of the residuals may be seen in the following table (11).

Table: The basic relationships between Variance, Covariance, Correlation and Residuals

	Variance	Covariance with X	Covariance with Y	Correlation with X	Correlation with Y
X	V_x	V_x	C_{xy}	1	r_{xy}
Y	V_y	C_{xy}	V_y	r_{xy}	1
$\hat{Y}$	$r_{xy}^2 V_y$	$C_{xy} = r_{xy}\sigma_x\sigma_y$	$r_{xy} V_y$	1	r_{xy}
$Y_r = Y - \hat{Y}$	$(1 - r_{xy}^2)V_y$	0	$(1 - r_{xy}^2)V_y$	0	$\sqrt{1 - r^2}$

Set theoretic approach: Partitioning the variance in Y





$$\beta_{y.x} = \frac{\sigma_{xy}}{\sigma_x^2}$$
$$\hat{y} = \beta_{y.x} X$$
$$r_{xy} = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2 \sigma_y^2}}$$
$$V_r = V_y + \frac{Cov_{xy}^2}{V_x} - 2\frac{Cov_{xy}^2}{V_x}$$
$$V_r = V_y - \frac{Cov_{xy}^2}{V_x}$$
$$V_r = V_y - r_{xy}^2 V_y$$
$$V_r = V_y(1 - r_{xy}^2)$$

Variance in Y predicted by $X = r_{xy}^2 \sigma_y^2$

Distance in the observational space

Because X and Y are vectors in the space defined by the observations, the covariance between them may be thought of in terms of the average squared distance between the two vectors in that same space. That is, following Pythagorus, the *distance*, d, is simply the square root of the sum of the squared distances in each dimension (for each pair of observations), or, if we find the average distance, we can find the square root of the sum of the squared distances divided by n:

$$d_{xy}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2.$$

which is the same as

$$d_{xy}^2 = V_x + V_y - 2C_{xy}$$

$$d_{xy} = \sqrt{2 * (1 - r_{xy})}. \tag{6}$$

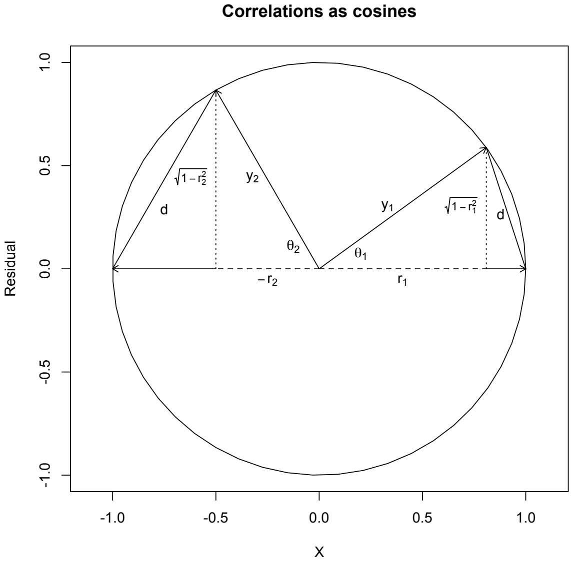
Distance, correlations, and the law of cosines

Compare this to the trigonometric *law of cosines*,

$$c^2 = a^2 + b^2 - 2ab \cdot cos(ab),$$

and we see that the distance between two vectors is the sum of their variances minus twice the product of their standard deviations times the *cosine* of the angle between them. That is, the correlation is the cosine of the angle between the two vectors. The next figure shows these relationships for two Y vectors. The correlation, r_1, of X with Y_1 is the cosine of $\theta_1 =$ the ratio of the projection of Y_1 onto X. From the *Pythagorean Theorem*, the length of the residual Y with X removed ($Y.x$) is $\sigma_y\sqrt{1 - r^2}$.

A geometric version of correlation and distance

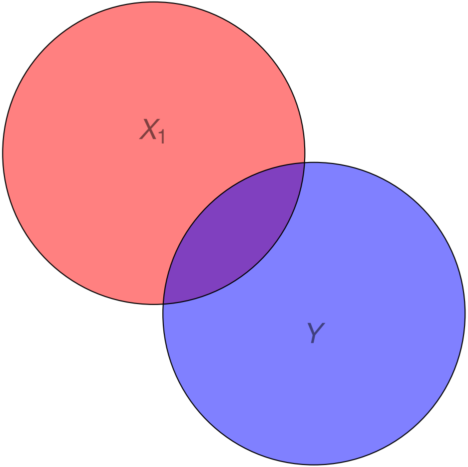


1. Projection of y on x is r
2. $\hat{y} = rx + \epsilon$
3. Residual (ϵ) is that part of y orthogonal to x
4. Residual (ϵ) $= \sqrt{1 - r^2}$
5. $cos(\theta) = r$
6. $d_{xy}^2 = (1-r)^2 + \sqrt{1 - r^2}^2$
7. $d_{xy}^2 = 1 - 2r + r^2 + (1 - r^2)$
8. $d_{xy} = 2(1 - r_{xy})$

R code

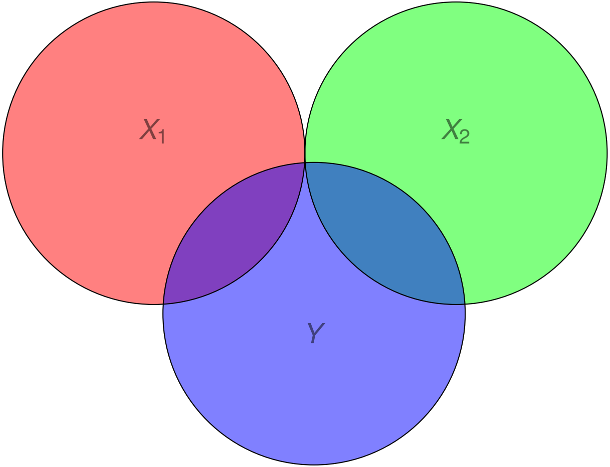
```
#Showing the law of cosines
 segments=51
 angles <- (0:segments) * 2 * pi/segments
        unit.circle <- cbind(cos(angles), sin(angles))
plot(unit.circle,typ="l",xlab="X",ylab="Residual",main="Correlations as cosines",asp=1)

theta <- pi/5   #the first correlation
x2 <- c(cos(theta),sin(theta))
segments(0,0,x2[1],0,lty="dashed")
arrows(x2[1],0,1,0,length=.075)
arrows(0,0,x2[1],x2[2],length=.075)
segments(x2[1],0,x2[1],x2[2],lty="dotted")
text(x2[1]/2,.0,expression(r[1]),pos=1)
text(x2[1],x2[2]/2,expression(sqrt(1-r[1]^2)),pos=2,cex=.8)
text(x2[1]/4,x2[2]/8,expression(theta[1]))
text(x2[1]/2,x2[2]/2, expression(y[1]),pos=2)
segments(x2[1],x2[2],1,0)    #show distance
text(x2[1]+.07,x2[2]*.45,'d')

theta <- 2*pi/3   #another correlation
x2 <- c(cos(theta),sin(theta))
segments(0,0,x2[1],0,lty="dashed")
arrows(x2[1],0,-1,0,length=.075)
arrows(0,0,x2[1],x2[2],length=.075)
segments(x2[1],0,x2[1],x2[2],lty="dotted")
text(x2[1]/2,.0,expression(-r[2]),pos=1)
text(x2[1],x2[2]/2,expression(sqrt(1-r[2]^2)),pos=2,cex=.8)
text(x2[1]/4,x2[2]/8,expression(theta[2]))
text(x2[1]/2,x2[2]/2, expression(y[2]),pos=2)
segments(-1,0,x2[1],x2[2])  #the distance
text(x2[1]*1.5,x2[2]/3,'d')
```

Venn diagram representation of predicting Y from X_1



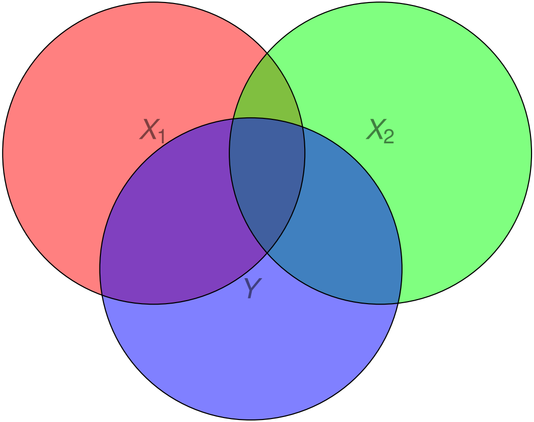
Variance in Y predicted by X_1 $\hat{V}_y = V_y r^2_{x_1 y}$

The Ideal model of predicting Y from X_1 and X_2



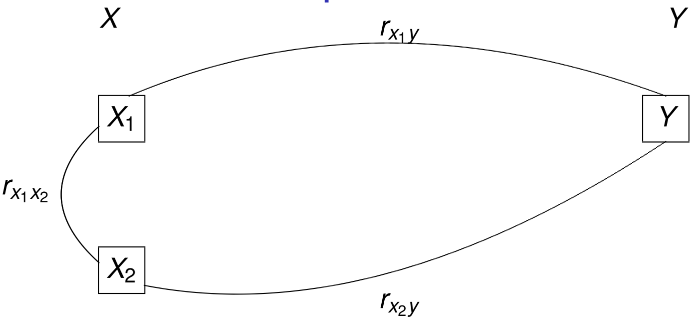
Variance in Y predicted by X_1 and X_2 if X_1 and X_2 are independent.
$\hat{V}_y = V_y r^2_{x_1 y} + V_y r^2_{x_2 y}$

The usual case of predicting Y from X_1 and X_2

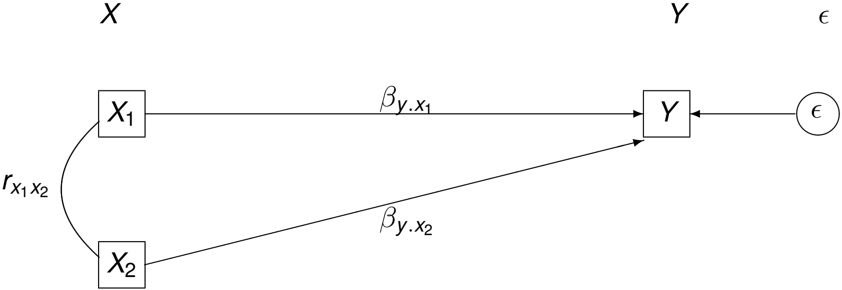


Variance in Y predicted by X_1 and X_2 if X_1 and X_2 - overlapping predictions $\hat{V}_y = V_y r^2_{x_1 y} + V_y r^2_{x_2 y} -$ overlap
But what is the overlap?

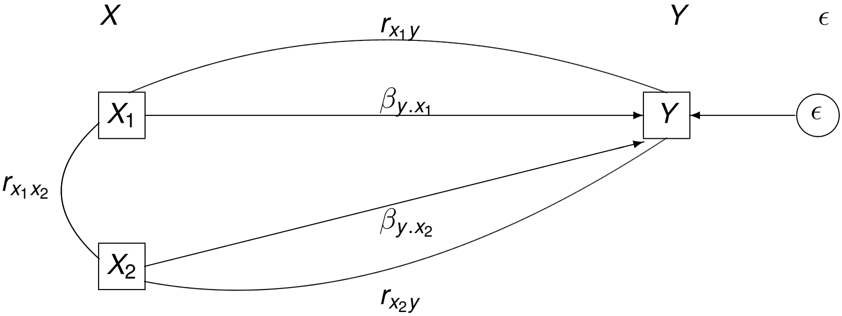
Multiple correlations



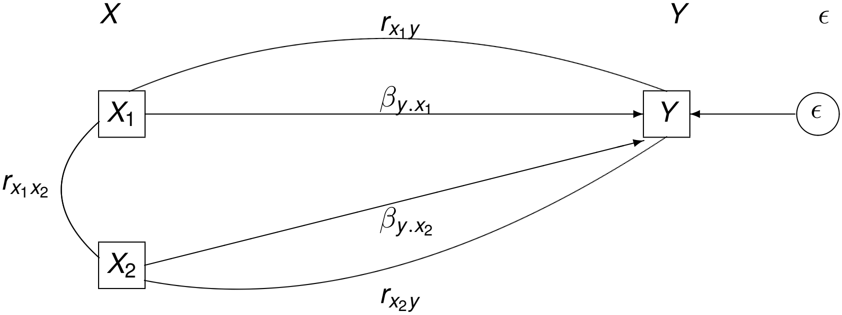
Multiple Regression



Multiple Regression: decomposing correlations



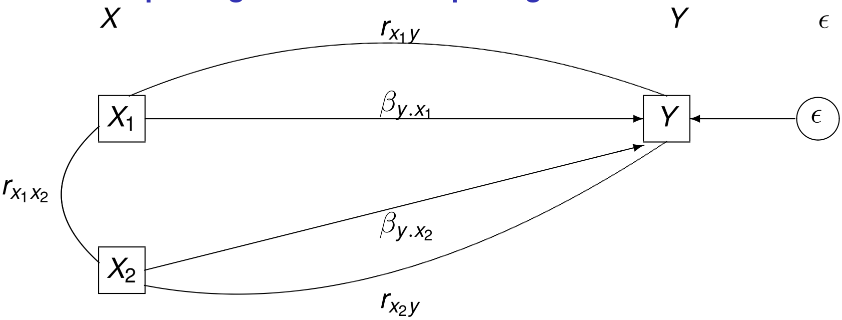
Multiple Regression: decomposing correlations



$$r_{x_1 y} = \overbrace{\beta_{y.x_1}}^{\text{direct}} + \overbrace{r_{x_1 x_2}\beta_{y.x_2}}^{\text{indirect}}$$

$$r_{x_2 y} = \underbrace{\beta_{y.x_2}}_{\text{direct}} + \underbrace{r_{x_1 x_2}\beta_{y.x_1}}_{\text{indirect}}$$

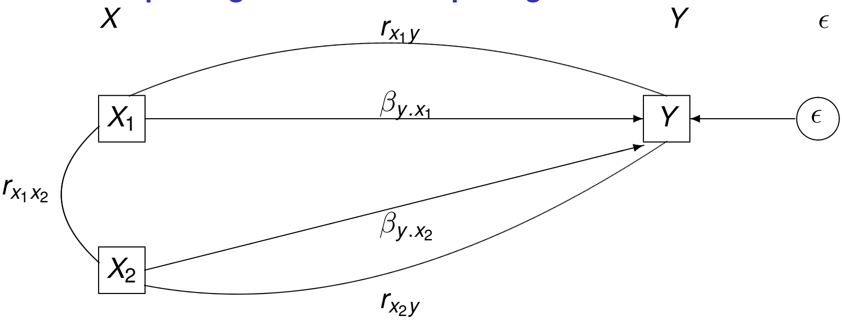
Multiple Regression: decomposing correlations



$$r_{x_1y} = \overbrace{\beta_{y.x_1}}^{direct} + \overbrace{r_{x_1x_2}\beta_{y.x_2}}^{indirect} \qquad \beta_{y.x_1} = \frac{r_{x_1y} - r_{x_1x_2}r_{x_2y}}{1 - r^2_{x_1x_2}}$$

$$r_{x_2y} = \underbrace{\beta_{y.x_2}}_{direct} + \underbrace{r_{x_1x_2}\beta_{y.x_1}}_{indirect} \qquad \beta_{y.x_2} = \frac{r_{x_2y} - r_{x_1x_2}r_{x_1y}}{1 - r^2_{x_1x_2}}$$

Multiple Regression: decomposing correlations



$$r_{x_1 y} = \overbrace{\beta_{y.x_1}}^{\text{direct}} + \overbrace{r_{x_1 x_2}\beta_{y.x_2}}^{\text{indirect}}$$

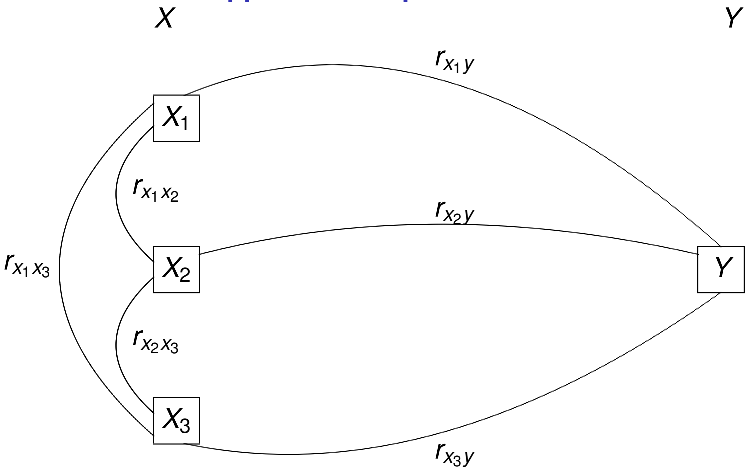
$$r_{x_2 y} = \underbrace{\beta_{y.x_2}}_{\text{direct}} + \underbrace{r_{x_1 x_2}\beta_{y.x_1}}_{\text{indirect}}$$

$$\beta_{y.x_1} = \frac{r_{x_1 y} - r_{x_1 x_2} r_{x_2 y}}{1 - r^2_{x_1 x_2}}$$

$$\beta_{y.x_2} = \frac{r_{x_2 y} - r_{x_1 x_2} r_{x_1 y}}{1 - r^2_{x_1 x_2}}$$

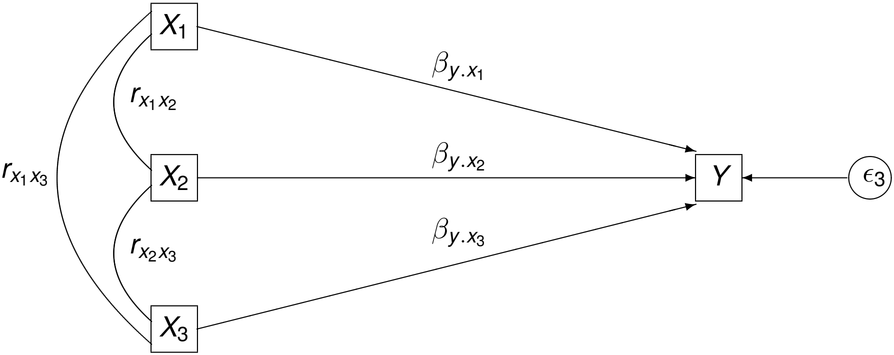
$$R^2 = r_{x_1 y}\beta_{y.x_1} + r_{x_2 y}\beta_{y.x_2}$$

What happens with 3 predictors? The correlations

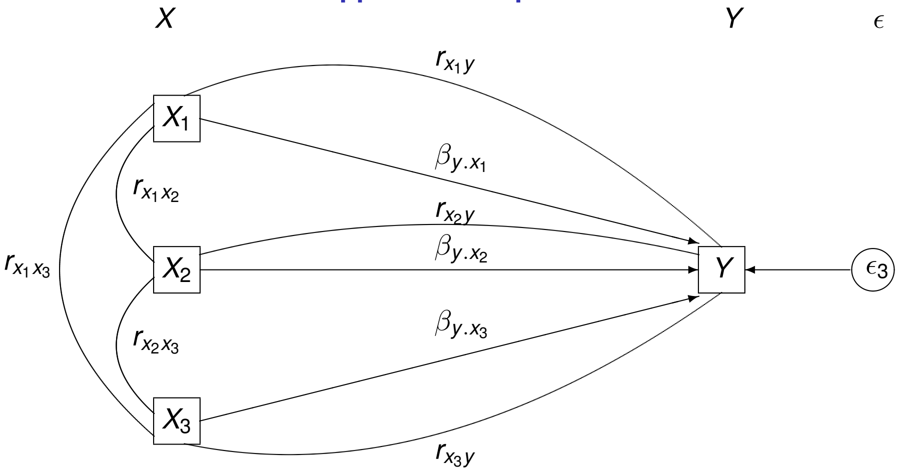


What happens with 3 predictors? β weights

X Y ϵ



What happens with 3 predictors?



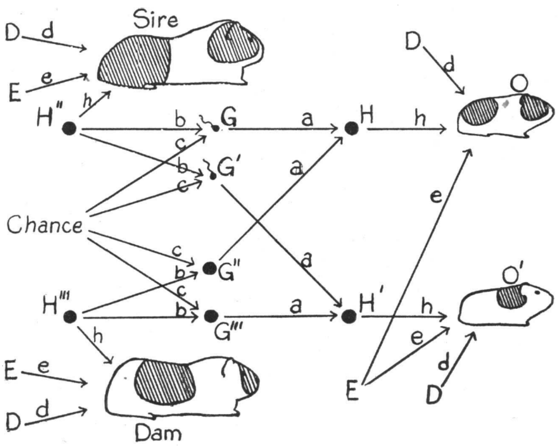
$$r_{x_1y} = \underbrace{\beta_{y.x_1}}_{direct} + \underbrace{r_{x_1x_2}\beta_{y.x_1} + r_{x_1x_3}\beta_{y.x_3}}_{indirect} \quad r_{x_2y} = \ldots \quad r_{x_3y} = \ldots$$

The math gets tedious

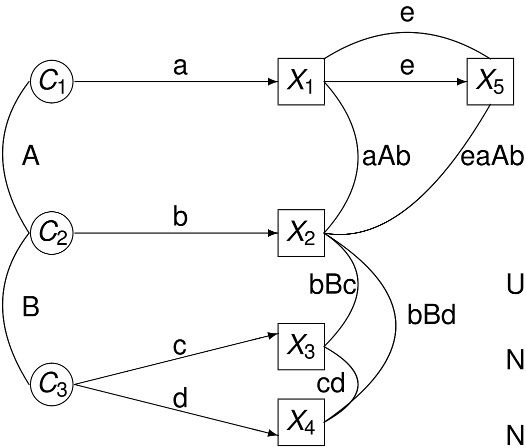
Multiple regression and linear algebra

- Multiple regression requires solving multiple, simultaneous equations to estimate the direct and indirect effects.
  - Each equation is expressed as a $r_{x_i y}$ in terms of direct and indirect effects.
  - Direct effect is $\beta_{y.x_i}$
  - Indirect effect is $\sum_{j \neq i} beta_{y.x_j} r_{x_j y}$
- How to solve these equations?
- Tediously, or just use linear algebra.

Wright's Path model of inheritance in the Guinea Pig (Wright, 1921)



The basic rules of path analysis–think genetics



Up ... and over and down ...

No down and up

No double overs

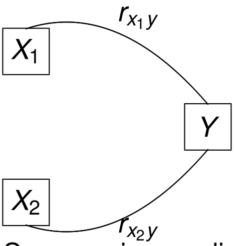
Up ... and down ...

Parents cause children

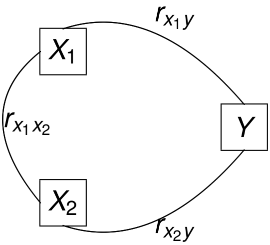
children do not cause parents

3 special cases of regression

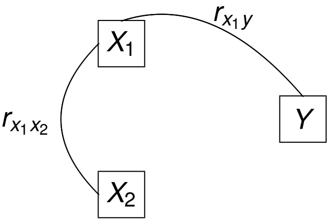
Orthogonal predictors



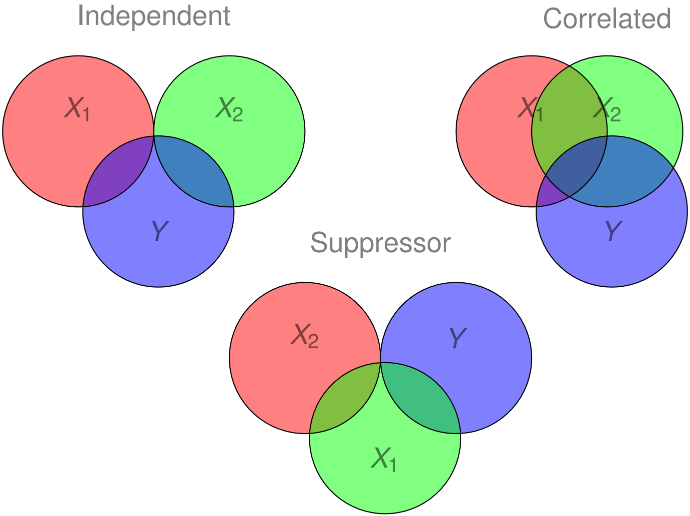
Correlated predictors



Suppressive predictors

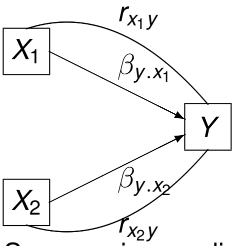


Three basic cases

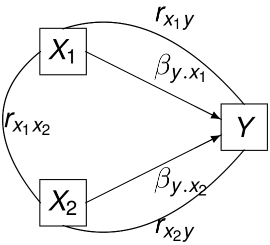


3 special cases of regression

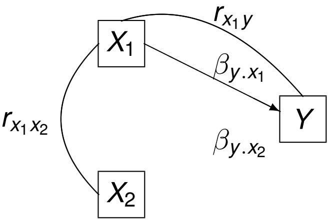
Orthogonal predictors



Correlated predictors



Suppressive predictors

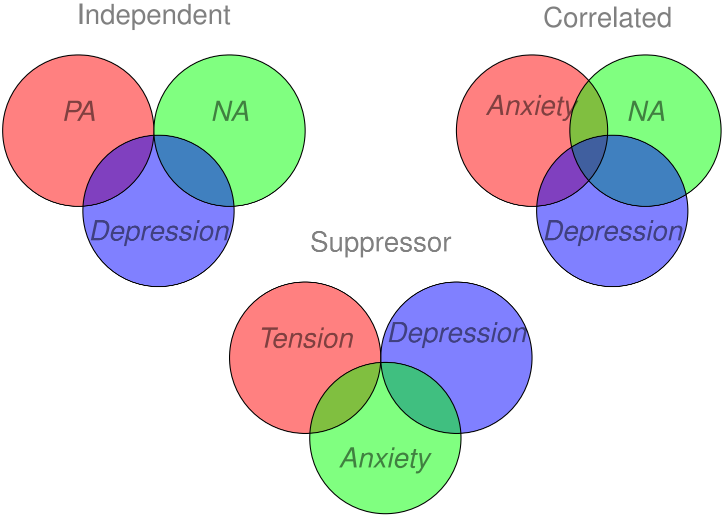


$$\beta_{y.x_1} = \frac{r_{x_1 y} - r_{x_1 x_2} r_{x_2 y}}{1 - r^2_{x_1 x_2}}$$

$$\beta_{y.x_2} = \frac{r_{x_2 y} - r_{x_1 x_2} r_{x_1 y}}{1 - r^2_{x_1 x_2}}$$

$$R^2 = r_{x_1 y}\beta_{y.x_1} + r_{x_2 y}\beta_{y.x_2}$$

Three basic cases: Theoretical examples



Sentence comprehension, age and grade

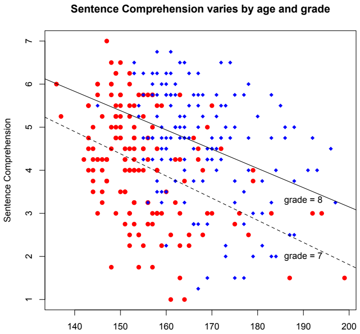
```
lowerCor(holzinger.swineford[cs(t07_sentcomp , agemo, grade)])

             t07_s agemo grade
t07_sentcomp  1.00
agemo        -0.23  1.00
grade         0.18  0.53  1.00
#plot the points
 plot(t07_sentcomp ~ agemo,
            col=c("red","blue")[holzinger.swineford$grade -6],
  pch=26-holzinger.swineford$grade,data=holzinger.swineford,
  ylab="Sentence Comprehension",xlab="Age in Months",
  main="Sentence Comprehension varies by age and grade")

 #add the lines
  by(holzinger.swineford, holzinger.swineford$grade -6,
          function(x) abline(
  lmCor(t07_sentcomp ~ agemo,data=x, std=FALSE, plot=FALSE)
     ,lty=c("dashed","solid")[x$grade-6]))
   #label them
text(190,3.3,"grade = 8")
 text(190,2,"grade = 7")
```

The regression

R code

```
lmCor(t07_sentcomp ~ agemo+grade,data=holzinger.swineford)
```

```
Multiple Regression from raw data
 DV =  t07_sentcomp
            slope   se     t       p lower.ci upper.ci  VIF Vy.x
(Intercept)  0.00 0.05  0.00 1.0e+00    -0.10     0.10 1.00 0.00
agemo       -0.46 0.06 -7.39 1.5e-12    -0.58    -0.34 1.39 0.11
grade        0.42 0.06  6.78 6.4e-11     0.30     0.54 1.39 0.07
 Multiple Regression
                R   R2   Ruw R2uw Shrunken R2 SE of R2 overall F df1 df2        p
t07_sentcomp 0.43 0.18 -0.03    0        0.18     0.04     32.97   2 298 1.16e-13
```



Compare to step wise or hierarchical

R code

```
mod1 <- lm(t07_sentcomp ~ agemo,data=holzinger.swineford)
mod2 <-  lm(grade ~ agemo,data=holzinger.swineford)
comp.p <-  predict(mod1)      #the predicted scores
grade.p <- predict(mod2)
comp.age <-  holzinger.swineford[,14] - comp.p  #the residuals
grade.age <- holzinger.swineford[,3] - grade.p
 #combine into 1 data
res.df <- data.frame(comp=holzinger.swineford[,14], age=holzinger.swi
comp.age=comp.age,   grade.age = grade.age,
 comp.pred=comp.p,
grade.pred=grade.p )
lowerCor(res.df)
```

```
lowerCor(res.df)
           comp  age   grade cmp.g grd.g cmp.p grd.p
comp        1.00
age        -0.23  1.00
grade       0.18  0.53  1.00
comp.age    0.97  0.00  0.31  1.00
grade.age   0.36  0.00  0.85  0.37  1.00
comp.pred   0.23 -1.00 -0.53  0.00  0.00  1.00
grade.pred -0.23  1.00  0.53  0.00  0.00 -1.00  1.00
```

Compare to step wise or hierarchical

R code

```
mod3 <- lm(t07_sentcomp ~ grade,data=holzinger.swineford)
mod4 <-  lm(agemo ~ grade,data=holzinger.swineford)
comp.p <-  predict(mod3)      #the predicted scores
age.pred <- predict(mod4)
comp.grade <-  holzinger.swineford[,14] - comp.p  #the residuals
age.grade<- holzinger.swineford[,7] - age.pred
 #combine into 1 data
res.df <- data.frame(comp=holzinger.swineford[,14], age=holzinger.swi
comp.age=comp.age,   grade.age = grade.age,
comp.grade = comp.grade , age.grade =age.grade,
 comp.pred=comp.p, grade.pred=grade.p )
lowerCor(res.df)
 lmCor(comp ~ age + grade.age,data=res.df)
  lmCor(comp ~ age + age.grade,data=res.df)
```

```
lowerCor(res.df)
           comp  age   grade comp.g grd.g cmp.gr ag.gr cmp.p grd.p
comp        1.00
age        -0.23  1.00
grade       0.18  0.53  1.00
comp.age    0.97  0.00  0.31  1.00
grade.age   0.36  0.00  0.85  0.37   1.00
comp.grade  0.98 -0.33  0.00  0.93   0.21  1.00
age.grade  -0.39  0.85  0.00 -0.19  -0.53 -0.39   1.00
comp.pred   0.18  0.53  1.00  0.31   0.85  0.00   0.00  1.00
grade.pred -0.23  1.00  0.53  0.00   0.00 -0.33   0.85  0.53  1.00
```

Compare the various models

R code

```
m1 <- lmCor(comp ~ age + grade,data=res.df)  #traditional (joint) mode
m2 <- lmCor(comp ~ age + grade.age, data=res.df)  #hierarchical 1
m3 <- lmCor(comp ~ age.grade + grade , data=res.df) #hierarchical 2
sum.df <-  data.frame(m1=m1$coefficients,m2=m2$coefficients,
     m3 = m3$coefficients)
 t.df <-  data.frame(m1=m1$t,m2=m2$t,m3 = m3$t)
 r.df <- data.frame(m1=m1$R,m2 = m2$R,m3 = m3$R)
 colnames(t.df ) <- colnames(sum.df) <- colnames(r.df)
round(rbind(sum.df,t.df,r.df),2)
m1
```

```
 round(rbind(sum.df,t.df),2)
                 m1    m2    m3
(Intercept)    0.00  0.00  0.00
age           -0.46 -0.23 -0.39
grade          0.42  0.36  0.18
(Intercept)1   0.00  0.00  0.00
age1          -7.39 -4.47 -7.39
grade1         6.78  6.78  3.37
comp           0.43  0.43  0.43
Call: lmCor(y = comp ~ age + grade, data = res.df)  Multiple Regression from raw data
 DV =  comp
            slope   se     t       p lower.ci upper.ci  VIF Vy.x
(Intercept)  0.00 0.05  0.00 1.0e+00    -0.10     0.10 1.00 0.00
age         -0.46 0.06 -7.39 1.5e-12    -0.58    -0.34 1.39 0.11
grade        0.42 0.06  6.78 6.4e-11     0.30     0.54 1.39 0.07
Residual Standard Error =  0.91  with  298  degrees of freedom
 Multiple Regression
        R   R2   Ruw R2uw Shrunken R2 SE of R2 overall F df1 df2       p
comp 0.43 0.18 -0.03    0        0.18     0.04     32.97   2 298 1.16e-13
```

Find the regression of rated Prelim score on GREV

```
> mod1 <- lm(GPA~GREV,data=mydata)
> summary(mod1)

Call:
lm(formula = GPA ~ GREV, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.45807 -0.32322  0.00107  0.32811  1.44850

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.0117292  0.0694343   43.38   <2e-16 ***
GREV        0.0019839  0.0001359   14.60   <2e-16 ***
---
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: 0.4558 on 998 degrees of freedom
Multiple R-squared: 0.176,      Adjusted R-squared: 0.1751
F-statistic: 213.1 on 1 and 998 DF,  p-value: < 2.2e-16
```

Regression on z transformed data

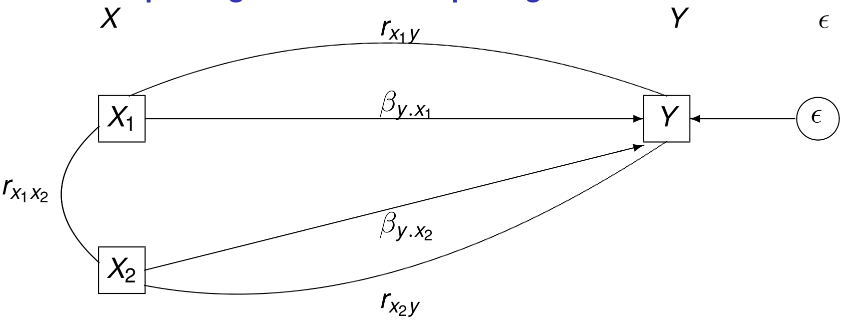
```
> mod2 <- lm(GPA~GREV,data=z.data)
> summary(mod2)

Call:
lm(formula = GPA ~ GREV, data = z.data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.90526 -0.64404  0.00213  0.65377  2.88619

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.888e-17  2.872e-02    0.00        1
GREV        4.195e-01  2.873e-02   14.60   <2e-16 ***
---
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: 0.9082 on 998 degrees of freedom
Multiple R-squared: 0.176,      Adjusted R-squared: 0.1751
F-statistic: 213.1 on 1 and 998 DF,  p-value: < 2.2e-16
```

Note that the slope is the same as the correlation.

```
> mod3 <- lm(GPA~GREV,data=cent)
> summary(mod3)

Call:
lm(formula = GPA ~ GREV, data = cent)

Residuals:
     Min       1Q   Median       3Q      Max
-1.45807 -0.32322  0.00107  0.32811  1.44850

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.332e-17  1.441e-02    0.00        1
GREV         1.984e-03  1.359e-04   14.60   <2e-16 ***
---
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: 0.4558 on 998 degrees of freedom
Multiple R-squared: 0.176,      Adjusted R-squared: 0.1751
F-statistic: 213.1 on 1 and 998 DF,  p-value: < 2.2e-16
```

Note that the slope of the centered data is in the same units as the raw data, just the intercept has changed.

Multiple Regression: decomposing correlations



$$r_{x_1 y} = \overbrace{\beta_{y.x_1}}^{\text{direct}} + \overbrace{r_{x_1 x_2}\beta_{y.x_2}}^{\text{indirect}}$$

$$r_{x_2 y} = \underbrace{\beta_{y.x_2}}_{\text{direct}} + \underbrace{r_{x_1 x_2}\beta_{y.x_1}}_{\text{indirect}}$$

$$\beta_{y.x_1} = \frac{r_{x_1 y} - r_{x_1 x_2} r_{x_2 y}}{1 - r^2_{x_1 x_2}}$$

$$\beta_{y.x_2} = \frac{r_{x_2 y} - r_{x_1 x_2} r_{x_1 y}}{1 - r^2_{x_1 x_2}}$$

$$R^2 = r_{x_1 y}\beta_{y.x_1} + r_{x_2 y}\beta_{y.x_2}$$

2 predictors

```
> summary(lm(GPA ~ GREV + GREQ , data= cent))

Call:
lm(formula = GPA ~ GREV + GREQ, data = cent)

Residuals:
     Min       1Q   Median       3Q      Max
-1.42442 -0.33228  0.00616  0.32465  1.43765

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.651e-17  1.435e-02   0.000  1.00000
GREV         1.534e-03  1.976e-04   7.760 2.10e-14 ***
GREQ         6.314e-04  2.019e-04   3.127  0.00182 **
---
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: 0.4538 on 997 degrees of freedom
Multiple R-squared: 0.184,       Adjusted R-squared: 0.1823
F-statistic: 112.4 on 2 and 997 DF,  p-value: < 2.2e-16
```

Multiple R with z transformed data

Do the same regression, but on the z transformed data. The units are now in correlation units.

```
> z.data <- data.frame(scale(my.data))
>  summary(lm(GPA ~ GREV + GREQ , data= z.data))

Call:
lm(formula = GPA ~ GREV + GREQ, data = z.data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.83821 -0.66208  0.01228  0.64688  2.86457

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.205e-17  2.860e-02   0.000  1.00000
GREV        3.242e-01  4.179e-02   7.760 2.10e-14 ***
GREQ        1.306e-01  4.179e-02   3.127  0.00182 **
---
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: 0.9043 on 997 degrees of freedom
Multiple R-squared: 0.184,       Adjusted R-squared: 0.1823
F-statistic: 112.4 on 2 and 997 DF,  p-value: < 2.2e-16
```

The 3 correlations produce the beta weights

```
> R.small <- cor(my.data[c(2,3,8)])
> round(R.small,2)
     GREV GREQ  GPA
GREV 1.00 0.73 0.42
GREQ 0.73 1.00 0.37
GPA  0.42 0.37 1.00
> solve(R.small[1:2,1:2])
          GREV      GREQ
GREV  2.133188 -1.554768
GREQ -1.554768  2.133188
> beta <- solve(R.small[1:2,1:2],
   R.small[3,1:2])
> beta
     GREV      GREQ
0.3242492 0.1306439
> beta.1 <- (.42 - .73*.37)/(1-.73^2)
> beta.1
[1] 0.3209163
> beta.2 <- (.37 - .73 * .42)/(1-.73^2)
> beta.2
[1] 0.1357311
```

- Find the correlation matrix
- Display it to two decimals
- Find the inverse of GREV and GREQ correlations
- Show them
- Find the beta weights by solving the matrix equation
- show them
- Find the beta weights by using the formula
- Show them

3 predictors, no interactions

Use three predictors, but print it with only 2 decimals

```
> print(summary(lm(GPA ~ GREV + GREQ + GREA , data= cent)),digits=3

Call:
lm(formula = GPA ~ GREV + GREQ + GREA, data = cent)

Residuals:
    Min       1Q   Median       3Q      Max
-1.2668 -0.3038  0.0073  0.3051  1.3022

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.89e-17   1.35e-02    0.00  1.00000
GREV         6.66e-04   2.00e-04    3.32  0.00092 ***
GREQ         7.75e-05   1.96e-04    0.40  0.69233
GREA         2.08e-03   1.81e-04   11.52  < 2e-16 ***
---
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: 0.427 on 996 degrees of freedom
Multiple R-squared: 0.28,       Adjusted R-squared: 0.278
F-statistic:  129 on 3 and 996 DF,  p-value: <2e-16
```

3 predictors, no interactions

Use three predictors, but just the middle 200 subjects

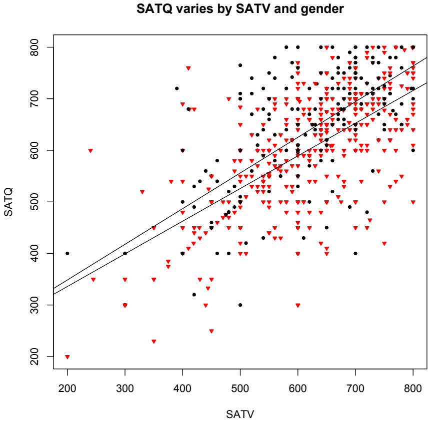
```
> mod4 <- lm(GPA ~ GREV + GREQ + GREA , data= cent[400:600,])
> summary(mod4)

Call:
lm(formula = GPA ~ GREV + GREQ + GREA, data = cent[400:600, ])

Residuals:
     Min       1Q   Median       3Q      Max
-1.03553 -0.30799 -0.00889  0.29320  1.20228

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0397399  0.0310412   1.280    0.202
GREV        0.0004706  0.0004530   1.039    0.300
GREQ        0.0005236  0.0004515   1.160    0.248
GREA        0.0017904  0.0004360   4.107 5.88e-05 ***
---
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: 0.4394 on 197 degrees of freedom
Multiple R-squared: 0.2259,      Adjusted R-squared: 0.2141
F-statistic: 19.16 on 3 and 197 DF,  p-value: 6.051e-11
```

Interaction terms are just products in regression

- To interpret all effects, the data need to be 0 centered.
  - This makes the main effects orthogonal to the interaction term.
  - Otherwise, need to compare model with and without interactions
- Graph the results in non-standardized form
- Consider a real data set of SAT V, SAT Q and Gender

```
> data(sat.act)
>  colors=c("black","red")          #choose some nice colors
>  symb=c(19,25)
>  colors=c("black","red")          #choose some nice colors
> with(sat.act,plot(SATQ~SATV,pch=symb[gender], col=colors[gender],
   bg=colors[gender],cex=.6,main="SATQ varies by SATV and gender"))
>  by(sat.act,sat.act$gender,function(x)
                  abline(lm(SATQ~SATV,data=x)))
```

An example of an interaction plot




```
>  data(sat.act)
>   c.sat <- data.frame(scale(sat
>   summary(lm(SATQ~SATV * gender

Call:
lm(formula = SATQ ~ SATV * gender

Residuals:
     Min        1Q   Median
-294.423  -49.876    5.577   53.2

Coefficients:
               Estimate Std. Error
(Intercept)  -0.26696    3.31211
SATV          0.65398    0.02926
gender      -36.71820    6.91495
SATV:gender  -0.05835    0.06086
---
Signif. codes:  0 Ô***Õ 0.001 Ô**

Residual standard error: 86.79 on
  (13 observations deleted due to
Multiple R-squared: 0.4391,
F-statistic: 178.3 on 3 and 683 D
```

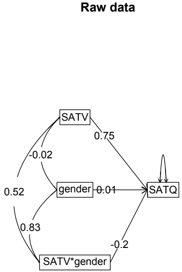
Interaction of Anxiety with Verbal

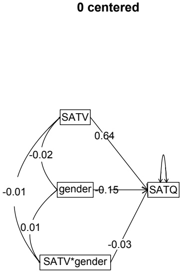
```
> mod5 <- lm(GPA ~ GREV * Anx, data=cent)
> summary(mod5)

Call:
lm(formula = GPA ~ GREV * Anx, data = cent)

Residuals:
     Min       1Q   Median       3Q      Max
-1.49677 -0.31527 -0.00054  0.31223  1.32156

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.375e-04  1.395e-02  -0.017    0.986
GREV         1.996e-03  1.316e-04  15.167  < 2e-16 ***
Anx         -1.131e-02  1.414e-03  -7.997 3.51e-15 ***
GREV:Anx     2.219e-05  1.377e-05   1.612    0.107
---


Residual standard error: 0.4412 on 996 degrees of freedom
Multiple R-squared: 0.2294,     Adjusted R-squared: 0.227
F-statistic: 98.81 on 3 and 996 DF,  p-value: < 2.2e-16
```

The effect of centering on interaction slopes

R code

```
mod0 <- lmCorr(SATQ ~ SATV *gender,std=TRUE,data=sat.act,zero=FALSE,ma

 mod1 <- lmCor(SATQ ~ SATV *gender,std=TRUE,data=sat.act,zero=TRUE,ma
```



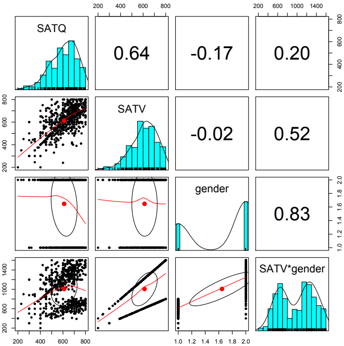


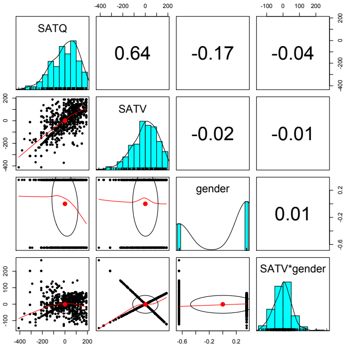
R code

```
mod0 <- lmCor(SATQ ~ SATV *gender,std=TRUE,data=sat.act,zero=FALSE,ma
mod1 <- lmCor(SATQ ~ SATV *gender,std=TRUE,data=sat.act,zero=TRUE,mai
```

```
Call: lmCor(y = SATQ ~ SATV * gender, data = sat.act, std = TRUE,
    main = "Raw data", zero = FALSE)

Multiple Regression from raw data

 DV =  SATQ
            slope   se     t        p   VIF
SATV         0.75 0.10  7.44 2.9e-13 12.68
gender       0.01 0.16  0.10 9.2e-01 30.32
SATV*gender -0.20 0.18 -1.10 2.7e-01 41.30

 Multiple Regression
        R   R2  Ruw R2uw Shrunken R2 SE of R2 overall F df1 df2        p
SATQ 0.67 0.44 0.65 0.43        0.44     0.03    184.19   3 696 6.69e-88
> mod1
Call: lmCor(y = SATQ ~ SATV * gender, data = sat.act, std = TRUE,
    main = "0 centered", zero = TRUE)

Multiple Regression from raw data

 DV =  SATQ
            slope   se     t        p VIF
SATV         0.64 0.03 22.72 6.4e-86   1
gender      -0.15 0.03 -5.41 8.5e-08   1
SATV*gender -0.03 0.03 -1.10 2.7e-01   1

 Multiple Regression
        R   R2  Ruw R2uw Shrunken R2 SE of R2 overall F df1 df2        p
SATQ 0.67 0.44 0.49 0.24        0.44     0.03    184.19   3 696 6.69e-88
```

Raw versus centered





Multiple R is just the optimal weighting of a set of variables

1. (Wilks, 1938) pointed out that as the number of items increases, differences between item weights become less important.
2. In the Robust Beauty of Improper Linear Models (Dawes, 1979), this property is suggested as showing that knowing the right variables to use is probably more important than knowing the precise weights.
3. Follows the principal of "it don't make no nevermind" (Wainer, 1976). That is, for standardized variables predicting a criterion with $.25 < \beta < .75$, setting all $beta_i = .5$ will reduce the accuracy of prediction by no more than 1/96th.
4. Thus the advice to standardize and add. (Clearly this advice does not work for strong negative correlations, but in that case standardize and subtract. In the general case weights of -1, 0, or 1 are the robust alternative.)
5. Also known as the concept of "fungible weights" (Waller, 2008).

Unit Weights versus optimal

Consider the Covariance Matrix: of XY

	X1	X2	...	Xn	Y
X1	V_{x_1}	$C_{x_1 x_2}$	...	$C_{x_1 x_n}$	$C_{x_1 y}$
X2	$C_{x_1 x_2}$	V_{x_2}	...	$C_{x_2 x_n}$	$C_{x_2 y}$
...	...	...	...	...	...
Xn	$C_{x_m x_1}$	$C_{x_n x_2}$	...	V_{x_n}	$C_{x_n y}$
Y	$C_{x_1 y}$	$C_{x_2 y}$	...	$C_{x_n y}$	V_y

$$\hat{Y} = \beta X \implies \beta = R^{-1} X$$

Multiply by β_{x_i} weights

	$\beta_{x_1} X1$	$\beta_{x_2} X2$	...	$\beta_{x_n} Xn$	Y
$\beta_{x_1} X1$	$\beta_{x_1} \beta_{x_1} V_{x_1}$	$\beta_{x_1} \beta_{x_2} C_{x_1 x_2}$	...	$\beta_{x_1} \beta_{x_n} C_{x_1 x_n}$	$C_{x_1 y}$
$\beta_{x_2} X2$	$\beta_{x_1} \beta_{x_2} C_{x_1 x_2}$	$\beta_{x_2} \beta_{x_2} V_{x_2}$	...	$\beta_{x_2} \beta_{x_n} C_{x_2 x_n}$	$C_{x_2 y}$
...	...	...	...	...	...
$\beta_{x_n} Xn$	$\beta_{x_1} \beta_{x_n} C_{x_m x_1}$	$\beta_{x_2} \beta_{x_n} C_{x_n x_2}$	...	$\beta_{x_n} \beta_{x_n} V_{x_n}$	$C_{x_n y}$
Y	$\beta_{x_1} C_{x_1 y}$	$\beta_{x_2} C_{x_2 y}$	...	$\beta_{x_n} C_{x_n y}$	V_y

Consider the example (but simulated) GRE achievement data

R code

```
datafilename=
"https://personality-project.org/r/datasets/psychometrics.prob2.txt"
mydata =read.file(datafilename)  #read the data file
lowerCor(mydata)
```

```
lowerCor(mydata)
       ID    GREV  GREQ  GREA  Ach   Anx   Prelm GPA   MA
ID      1.00
GREV   -0.01  1.00
GREQ    0.00  0.73  1.00
GREA   -0.01  0.64  0.60  1.00
Ach     0.00  0.01  0.01  0.45  1.00
Anx    -0.01  0.01  0.01 -0.39 -0.56  1.00
Prelim  0.02  0.43  0.38  0.57  0.30 -0.23  1.00
GPA     0.00  0.42  0.37  0.52  0.28 -0.22  0.42  1.00
MA     -0.01  0.32  0.29  0.45  0.26 -0.22  0.36  0.31  1.00
```

Predict GPA optimally using GREV + GREQ, versus unit weight them

R code
```
lmCor(GPA  ~ GREV + GREQ ,data=mydata)
```

```
Call: lmCor(y = GPA ~ GREV + GREQ, data = mydata)

Multiple Regression from raw data

 DV =  GPA
 intercept =  -223.44
     slope   se    t       p lower.ci upper.ci  VIF
GREV  0.32 0.04 7.76 2.1e-14     0.24     0.41 2.13
GREQ  0.13 0.04 3.13 1.8e-03     0.05     0.21 2.13

 Multiple Regression
       R   R2  Ruw R2uw Shrunken R2 SE of R2 overall F df1 df2        p SE residual
GPA 0.43 0.18 0.42 0.18        0.18     0.02    112.37   2 997 9.76e-45         0.9
```

Note that the R^2 goes from .18 to .18 even though we are weighting them equally versus 2.5 times as much!

Compare various weightings

R code

```
optimal <- .32 * mydata$GREV + .13* mydata$GREQ
#Correlated optimal weighting with criterion
suboptimal <- .13 * mydata$GREV + .32* mydata$GREQ
equal <-  mydata$GREV +  mydata$GREQ
lowerCor(example[cs(GREV,GREQ, GPA,optimal,suboptimal,equal)])
```

```
           GREV GREQ GPA  optml sbptm equal
GREV        1.00
GREQ        0.73 1.00
GPA         0.42 0.37 1.00
optimal     0.98 0.85 0.43 1.00
suboptimal  0.86 0.98 0.41 0.95  1.00
equal       0.93 0.93 0.42 0.99  0.99  1.00
```

Add in more predictors

R code

```
lmCor(MA ~ GREV + GREQ + GREA + Ach + Prelim + GPA + Anx,data=mydata)
```

```
Call: lmCor(y = MA ~ GREV + GREQ + GREA + Ach + Prelim + GPA + Anx,
    data = mydata)

Multiple Regression from raw data

 DV =  MA
 intercept =  -168.33
       slope   se     t       p lower.ci upper.ci  VIF
GREV    0.08 0.05  1.61 1.1e-01    -0.02     0.17 2.82
GREQ    0.02 0.04  0.41 6.8e-01    -0.07     0.10 2.37
GREA    0.24 0.05  4.74 2.4e-06     0.14     0.35 3.44
Ach     0.08 0.04  2.00 4.6e-02     0.00     0.15 1.83
Prelim  0.12 0.03  3.49 5.0e-04     0.05     0.19 1.56
GPA     0.06 0.03  1.82 6.9e-02     0.00     0.13 1.45
Anx    -0.04 0.04 -1.19 2.3e-01    -0.11     0.03 1.59

 Multiple Regression
      R   R2  Ruw R2uw Shrunken R2 SE of R2 overall F df1 df2        p SE residual
MA 0.48 0.23 0.47 0.22        0.23     0.02     42.71   7 992 7.94e-53        0.88
```

Unit weighted scores are simply sum scores of standardized variables.

Using `lmCor` and `mediate` for regressions

- lmCor in the `psych` package does multiple regressions (with or without interactions) from the correlation matrix or from the raw data.
- Mediate will do mediation analysis
- But, lmCor will do several multiple regressions at the same time.
- Also, lmCor will find the correlation between the predictor set of variables and the criterion set of variables.

lmCor

Using our data set, first find the correlations. Then show the correlations to two decimals using the `lower.mat` function.

```
> my.R <- lowerCor(mydata) #combines cor and loweMat
       ID     GREV  GREQ  GREA  Ach   Anx   Prelm GPA   MA
ID      1.00
GREV   -0.01  1.00
GREQ    0.00  0.73  1.00
GREA   -0.01  0.64  0.60  1.00
Ach     0.00  0.01  0.01  0.45  1.00
Anx    -0.01  0.01  0.01 -0.39 -0.56  1.00
Prelim  0.02  0.43  0.38  0.57  0.30 -0.23  1.00
GPA     0.00  0.42  0.37  0.52  0.28 -0.22  0.42  1.00
MA     -0.01  0.32  0.29  0.45  0.26 -0.22  0.36  0.31  1.00
```

Now, find the multiple regression of the first five (not counting ID) variables and the last three. This is in some sense snooping the data.

lmCor: regressions from covariance matrices

First, find the correlations, then do the regression

```
> my.R <- cor(mydata)
>  lmCor(y=c(7:9),x=2:6,data=my.R) #old way
#or
 lmCor(Prelim + GPA + MA ~ GREV + GREQ + GREA + Ach+ Anx,data=my.R)

Call: lmCor(y = Prelim + GPA + MA ~ GREV + GREQ + GREA + Ach + Anx,
    data = my.R)

Multiple Regression from matrix input

Beta weights
      Prelim   GPA    MA
GREV    0.14  0.20  0.10
GREQ    0.04  0.05  0.03
GREA    0.40  0.29  0.31
Ach     0.11  0.12  0.10
Anx    -0.01 -0.05 -0.05
Multiple R
Prelim    GPA     MA
  0.59   0.54   0.47
Multiple R2
Prelim    GPA     MA
  0.34   0.29   0.22
```

lmCor (for matrix based regressions)

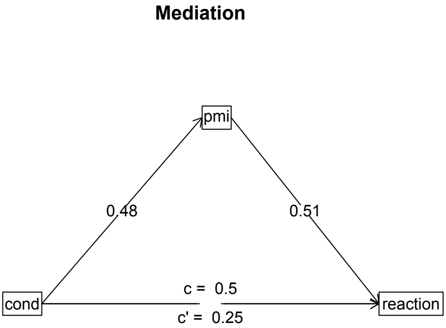
Specifying the number of observations gives significance tests.

```
> set.cor(data=my.R,x=c(2:6),y=c(7:9),n.obs=1000)
Call: set.cor(y = c(7:9), x = c(2:6), data = my.R, n.obs = 1000)
Multiple Regression from matrix input
Beta weights
     Prelim  GPA   MA
GREV   0.14 0.20 0.10
GREQ   0.04 0.05 0.03
...
Multiple R
Prelim    GPA    MA
  0.59   0.54  0.47
Multiple R2
Prelim    GPA    MA
  0.34   0.29  0.22
 SE of Beta weights
     Prelim GPA  MA
GREV   0.04 0.04 0.05
...
 t of Beta Weights
     Prelim  GPA   MA
GREV   3.28 4.50 2.24
...
Probability of t <
      Prelim     GPA      MA
...
 Shrunken R2
Prelim    GPA    MA
  0.34   0.29  0.21
Standard Error of R2
Prelim    GPA    MA
 0.024  0.024  0.023
```

Mediation is a special multiple regression model

1. “Tal-Or et al. (2010) examined the presumed effect of the media in two experimental studies. These data are from study 2. ’... perceptions regarding the influence of a news story about an expected shortage in sugar were manipulated indirectly, by manipulating the perceived exposure to the news story, and behavioral intentions resulting from the story were consequently measured." (p 801)."
2. IV is news story
3. DV is behavioral intentions
4. Effect is thought to be *mediated* through Perceived Media Exposure
5. IV -> DV (c path is the direct effect) item IV - > Mediator (a path)
6. Mediator -> DV (b path) item ab is indirect path, c’ is c - ab

Mediation in the Tal Or experiment



R code

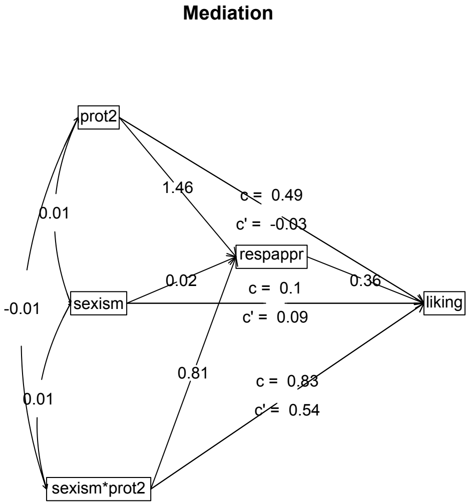
```
mediate(reaction ~ cond + (pmi), data =Tal_Or,n.iter=50)
```

```
Mediation/Moderation Analysis
Call: mediate(y = reaction ~ cond + (pmi), data = Tal_Or, n.iter = 50)

The DV (Y) was  reaction . The IV (X) was  cond . The mediating variable(s) =  pmi .

Total effect(c) of  cond  on  reaction  =  0.5   S.E. =  0.28  t  =  1.79  df=  120   with p
Direct effect (c') of  cond  on  reaction  removing  pmi  =  0.25   S.E. =  0.26  t  =  0.99
Indirect effect (ab) of  cond  on  reaction  through  pmi   =  0.24
Mean bootstrapped indirect effect =  0.24  with standard error =  0.14  Lower CI =  0.01    U
R = 0.45 R2 = 0.21   F = 15.56 on 2 and 120 DF   p-value:  9.83e-07

 To see the longer output, specify short = FALSE in the print statement or ask for the summar
```

Moderated mediation

1. “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")."
2. The interactive effect of IV on DV is mediated by M
3. Need to find the product terms

Moderated mediation graphiically



R code

```
mediate(liking ~  sexism * prot2 + (respappr), data=Garcia, n.iter = 5
```

```
Mediation/Moderation Analysis
Call: mediate(y = liking ~ sexism * prot2 + (respappr), data = Garcia,
    n.iter = 50)

The DV (Y) was  liking . The IV (X) was  sexism prot2 sexism*prot2 . The mediating variable(s

Total effect(c) of  sexism  on  liking  =  0.1   S.E. =  0.11  t  =  0.86  df=  124   with p
Direct effect (c') of  sexism  on  liking  removing  respappr  =  0.09   S.E. =  0.1  t  =  0
Indirect effect (ab) of  sexism  on  liking  through  respappr   =  0.01
Mean bootstrapped indirect effect =  0.01  with standard error =  0.05  Lower CI =  -0.08

Total effect(c) of  prot2  on  liking  =  0.49   S.E. =  0.19  t  =  2.63  df=  124   with p
Direct effect (c') of  prot2  on  NA  removing  respappr  =  -0.03   S.E. =  0.2  t  =  -0.16
Indirect effect (ab) of  prot2  on  liking  through  respappr   =  0.52
Mean bootstrapped indirect effect =  0.01  with standard error =  0.05  Lower CI =  0.32    U

Total effect(c) of  sexism*prot2  on  liking  =  0.83   S.E. =  0.24  t  =  3.42  df=  124
Direct effect (c') of  sexism*prot2  on  NA  removing  respappr  =  0.54   S.E. =  0.23  t  =
Indirect effect (ab) of  sexism*prot2  on  liking  through  respappr   =  0.29
Mean bootstrapped indirect effect =  0.01  with standard error =  0.05  Lower CI =  0.14    U
R = 0.53 R2 = 0.28   F = 12.26 on 4 and 124 DF   p-value:  1.99e-08

 To see the longer output, specify short = FALSE in the print statement or ask for the summar
```

Partial Correlation

1. Remove the effect of a z variable from the relationship between X and Y
   - Can show this for a single triple of variables or
   - As a matrix equation
2. $$r_{(x_i.x_j)(y.x_j)} = \frac{r_{x_i y} - r_{x_i x_j} r_{x_j y}}{\sqrt{(1 - r^2_{x_i x_j})(1 - r^2_{y x_j})}} \quad (7)$$
3. $\boldsymbol{X}^* = \boldsymbol{X} - \boldsymbol{R}_{xz}\boldsymbol{R}_z^{-1}\boldsymbol{Z}$
4. $\boldsymbol{C}^* = (\boldsymbol{R} - \boldsymbol{R}_{xz}\boldsymbol{R}_z^{-1})$
5. $\boldsymbol{R}^* = (\sqrt{diag(\boldsymbol{C}^*)}^{-1}\boldsymbol{C}^*\sqrt{diag(\boldsymbol{C}^*)}^{-1}$

Consider the following correlation matrix of Extraversion, 2 aspects of extraversion, and 4 measures of mood

josh

	b5.EXT	b5.EASS	b5.EENT	swb.tot	i.MP.PA	i.SWL	i.moodreg
b5.EXT	1.00	0.89	0.88	0.59	0.65	0.35	0.50
b5.EASS	0.89	1.00	0.55	0.40	0.58	0.25	0.35
b5.EENT	0.88	0.55	1.00	0.65	0.56	0.38	0.54
swb.tot	0.59	0.40	0.65	1.00	0.55	0.46	0.62
i.MP.PA	0.65	0.58	0.56	0.55	1.00	0.53	0.56
i.SWL	0.35	0.25	0.38	0.46	0.53	1.00	0.48
i.moodreg	0.50	0.35	0.54	0.62	0.56	0.48	1.00

What is the relationship of the mood measures when removing extraversion

```
> partial.r(m=josh,x=4:7,y=1)

partial correlations
           swb.tot i.MP.PA i.SWL i.moodreg
swb.tot       1.00    0.27  0.34      0.46
i.MP.PA       0.27    1.00  0.42      0.36
i.SWL         0.34    0.42  1.00      0.38
i.moodreg     0.46    0.36  0.38      1.00
```

Compare removing Assertiveness versus Enthusiasm

```
> partial.r(m=josh,x=4:7,y=3)
> partial.r(m=josh,x=4:7,y=2)
```

```
partial correlations
          swb.tot i.MP.PA i.SWL i.moodreg
swb.tot      1.00    0.30  0.30      0.42
i.MP.PA      0.30    1.00  0.41      0.37
i.SWL        0.30    0.41  1.00      0.35
i.moodreg    0.42    0.37  0.35      1.00

partial correlations
          swb.tot i.MP.PA i.SWL i.moodreg
swb.tot      1.00    0.43  0.41      0.56
i.MP.PA      0.43    1.00  0.49      0.47
i.SWL        0.41    0.49  1.00      0.43
i.moodreg    0.56    0.47  0.43      1.00
```

Original versus a partialed matrix

R code

```
lower <- lowerCor(mydata[-1])
upper <- partial.r(mydata[-1])
Rlow.up <- lowerUpper(lower,upper)
round(Rlow.up,2)
```

```
round(Rlow.up,2)
       GREV GREQ  GREA   Ach   Anx Prelim   GPA    MA
GREV     NA 0.45  0.39 -0.22  0.16   0.08  0.12  0.05
GREQ   0.73   NA  0.28 -0.14  0.09   0.02  0.03  0.01
GREA   0.64 0.60    NA  0.36 -0.26   0.22  0.13  0.15
Ach    0.01 0.01  0.45    NA -0.34   0.08  0.09  0.06
Anx    0.01 0.01 -0.39 -0.56    NA   0.00 -0.04 -0.04
Prelim 0.43 0.38  0.57  0.30 -0.23     NA  0.15  0.11
GPA    0.42 0.37  0.52  0.28 -0.22   0.42    NA  0.06
MA     0.32 0.29  0.45  0.26 -0.22   0.36  0.31    NA
```

Note how the sign of the partial correlation can be different from the raw correlation.

But, if we drop some of the predictors, the others seem important

R code

```
lower <- lowerCor(mydata[-c(1,2,4,5)])
upper <- partial.r(mydata[-c(1,2,4,5)])
Rlow.up <- lowerUpper(lower,upper)
round(Rlow.up,2)
```

```
       GREQ   Anx Prelim   GPA    MA
GREQ     NA  0.16   0.25  0.24  0.16
Anx    0.01    NA  -0.15 -0.15 -0.15
Prelim 0.38 -0.23     NA  0.25  0.20
GPA    0.37 -0.22   0.42    NA  0.12
MA     0.29 -0.22   0.36  0.31    NA
```

Which to drop? Try GREQ

```
      GREV   Anx Prelim   GPA    MA
GREV     NA  0.20   0.29  0.29  0.18
Anx    0.01    NA  -0.16 -0.17 -0.16
Prelim 0.43 -0.23     NA  0.22  0.18
GPA    0.42 -0.22   0.42    NA  0.10
MA     0.32 -0.22   0.36  0.31    NA
```

Testing for the significance of correlations

```
> 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 values (Entries above the diagonal are adjusted for multiple t
          gender education  age  ACT SATV SATQ
gender      0.00      0.17 1.00 1.00    1    0
education   0.02      0.00 0.00 0.00    1    1
```

Various tests of significance

1. Is the correlation different from 0? `cor.test`, `corr.test` (for more than two variables)
2. Does a correlation differ from another correlation, `r.test` with or without a third variable.
3. Does a correlation matrix differ from an Identity matrix? `cortest`
4. Bootstrapping confidence intervals for correlations `cor.ci`

Multiple R, Squared Multiple R, colinearity

1. When finding multiple R to predict one variable, we are finding the inverse of the $\boldsymbol{R}$ matrix ($\boldsymbol{R}^{-1}$) the diagonal of which is the residual variance of a variable when all others are removed.
2. Thus, the Squared Multiple R (SMC) of each variable is just $1 - \frac{1}{(1 - diag(\boldsymbol{R}^{-1})}$

```
round(smc(sat.act),2)
 gender education       age       ACT      SATV      SATQ
  0.06      0.32      0.32      0.43      0.47      0.51
```

3. The “Multiple Inflation Factor" is sometimes used as an index of colinearity and is $\frac{1}{1 - smc}$ which is the same as $diag(\boldsymbol{R}^{-1})$

```
vif <- 1/(1-smc(sat.act))
round(vif,2)
 gender education       age       ACT      SATV      SATQ
 1.06      1.47      1.47      1.74      1.90      2.05
# or
round(diag(solve(R)),2)
gender education       age       ACT      SATV      SATQ
 1.06      1.47      1.47      1.74      1.90      2.05
```

Mediation and moderation are sometimes used to explore “causal" links in regression models

1. Direct effect of X on Y (c)
2. Direct effect of X on M (a)
3. Direct effect of M on Y (b)
4. “Indirect Effect" of X on Y through M (ab)
5. Compare c to c - ab

However, just because you can specify a “causal” regression model, does not make it so.

The “Sobel" example from Preacher and Hayes (2004)

R code

```
?mediate  #produces this correlation matrix
sobel <- structure(list(SATIS = c(-0.59, 1.3, 0.02, 0.01, 0.79, -0.35,
-0.03, 1.75, -0.8, -1.2, -1.27, 0.7, -1.59, 0.68, -0.39, 1.33,
-1.59, 1.34, 0.1, 0.05, 0.66, 0.56, 0.85, 0.88, 0.14, -0.72,
0.84, -1.13, -0.13, 0.2), THERAPY = structure(c(0, 1, 1, 0, 1,
1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0, 1,
1, 1, 1, 0), value.labels = structure(c(1, 0), .Names = c("cognitive",
"standard"))), ATTRIB = c(-1.17, 0.04, 0.58, -0.23, 0.62, -0.26,
-0.28, 0.52, 0.34, -0.09, -1.09, 1.05, -1.84, -0.95, 0.15, 0.07,
-0.1, 2.35, 0.75, 0.49, 0.67, 1.21, 0.31, 1.97, -0.94, 0.11,
-0.54, -0.23, 0.05, -1.07)), .Names = c("SATIS", "THERAPY", "ATTRIB"
), row.names = c(NA, -30L), class = "data.frame", variable.labels = structure(c("Satisfaction"
"Therapy", "Attributional Positivity"), .Names = c("SATIS", "THERAPY",
"ATTRIB")))
R <- lowerCor(sobel)
lmCor(y="SATIS",x= c("ATTRIB","THERAPY"), data=sobel)
```

```
        SATIS THERA ATTRI
SATIS   1.00
THERAPY 0.43  1.00
ATTRIB  0.51  0.46  1.00
```

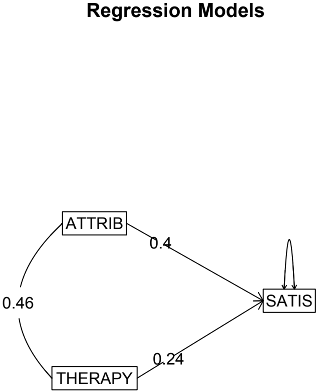
Try this as a mediation model (doing as a standardized regression

R code

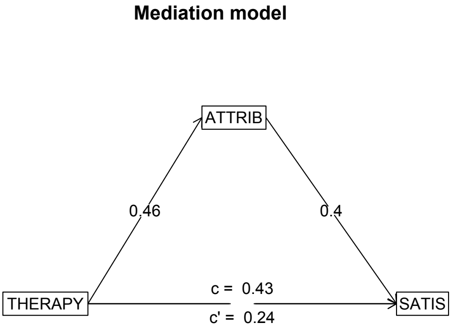
```
mediate(y="SATIS", x = "THERAPY", m="ATTRIB", data=sobel,std=TRUE)
```

```
The DV (Y) was  SATIS . The IV (X) was  THERAPY . The mediating variable(s) =  ATTRIB .

Total Direct effect(c) of  THERAPY  on  SATIS  =  0.43   S.E. =  0.17  t direct =  2.5   with
Direct effect (c') of  THERAPY  on  SATIS  removing  ATTRIB  =  0.24   S.E. =  0.18  t direct
Indirect effect (ab) of  THERAPY  on  SATIS  through  ATTRIB   =  0.18
Mean bootstrapped indirect effect =  0.18  with standard error =  0.09  Lower CI =  0.02    U
R2 of model =  0.31
 To see the longer output, specify short = FALSE in the print statement

 Total effect estimates (c)
        SATIS   se    t   Prob
THERAPY  0.43 0.17 2.5 0.0186

Direct effect estimates     (c')
        SATIS   se     t  Prob
THERAPY  0.24 0.18 1.35 0.190
ATTRIB   0.40 0.18 2.23 0.034

 'a'  effect estimates
       THERAPY   se    t   Prob
ATTRIB    0.46 0.17 2.74 0.0106

 'b'  effect estimates
       SATIS   se    t  Prob
ATTRIB   0.4 0.18 2.23 0.034

 'ab'  effect estimates
        SATIS boot   sd lower upper
THERAPY  0.18 0.18 0.09  0.02  0.38
>
```

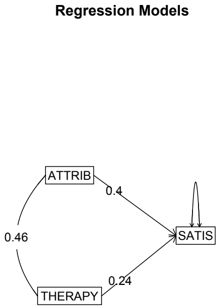
The simple path model of the sobel data set

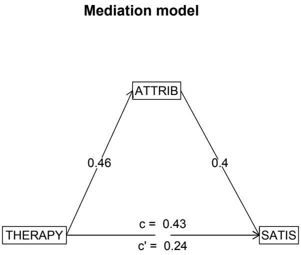


Mediation (standardized coefficients)



Compare regression to Mediation (standardized coefficients)





Input a covariance matrix into mediate

R code

```
C.pmi <- structure(c(0.251232840197254, 0.119718779155005, 0.157470345195255,
+ 0.124533519925363, 0.03052112488338, 0.0734039717446355, 0.119718779155005,
+ 1.74573503931761, 0.647207783553245, 0.914575836332134, 0.0133613221378115,
+ -0.0379181660669066, 0.157470345195255, 0.647207783553245, 3.01572704251633,
+ 1.2512828202525, -0.0224576835932294, 0.73973743835799, 0.124533519925363,
+ 0.914575836332134, 1.2512828202525, 2.40342196454751, -0.0106624017059843,
+ -0.752990470478475, 0.03052112488338, 0.0133613221378115, -0.0224576835932294,
+ -0.0106624017059843, 0.229241636678662, 0.884479541516727, 0.0734039717446355,
+ -0.0379181660669066, 0.73973743835799, -0.752990470478475, 0.884479541516727,
+ 33.6509729441557), .Dim = c(6L, 6L), .Dimnames = list(c("cond",
+ "pmi", "import", "reaction", "gender", "age"), c("cond", "pmi",
+ "import", "reaction", "gender", "age")))
```

```
R <- lowerMat(C.pmi)
         cond  pmi   imprt rectn gendr age
cond      0.25
pmi       0.12  1.75
import    0.16  0.65  3.02
reaction  0.12  0.91  1.25  2.40
gender    0.03  0.01 -0.02 -0.01  0.23
age       0.07 -0.04  0.74 -0.75  0.88 33.65
```

Mediation model

R code

```
mediate(y="reaction",x = "cond",m=c("pmi","import"),data=C.pmi,n.obs=
```

```
The DV (Y) was  reaction . The IV (X) was  cond . The mediating variable(s) =  pmi import .

Total Direct effect(c) of  cond  on  reaction  =  0.5   S.E. =  0.28  t direct =  1.79   with
Direct effect (c') of  cond  on  reaction  removing  pmi import  =  0.1   S.E. =  0.24  t dir
Indirect effect (ab) of  cond  on  reaction  through  pmi import   =  0.39
Mean bootstrapped indirect effect =  0.34  with standard error =  0.17  Lower CI =  0.01    U
R2 of model =  0.33
 To see the longer output, specify short = FALSE in the print statement
 Total effect estimates (c)
     reaction   se    t   Prob
cond      0.5 0.28 1.79 0.0766
Direct effect estimates     (c')
       reaction   se    t     Prob
cond       0.10 0.24 0.43 6.66e-01
pmi        0.40 0.09 4.26 4.04e-05
import     0.32 0.07 4.59 1.13e-05
 'a'  effect estimates
       cond   se    t   Prob
pmi    0.48 0.24 2.02 0.0452
import 0.63 0.31 2.02 0.0452
 'b'  effect estimates
       reaction   se    t     Prob
pmi        0.40 0.09 4.26 4.04e-05
import     0.32 0.07 4.59 1.13e-05
 'ab'  effect estimates
     reaction boot   sd lower upper
cond     0.39 0.34 0.17  0.01  0.72
```

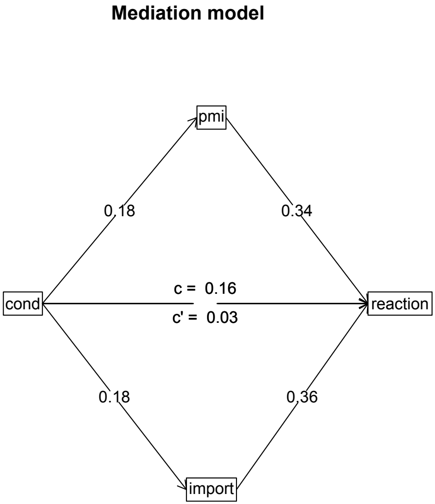
The Covariance matrix and the correlation matrix

R code

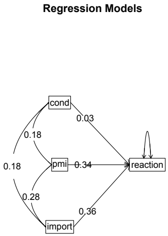
```
lowerMat(C.pmi)
lowerMat(cov2cor(C.pmi))
```

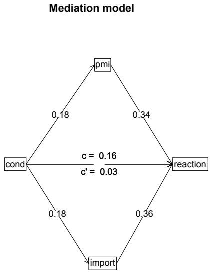
```
         cond  pmi   imprt rectn gendr age
cond      0.25
pmi       0.12  1.75
import    0.16  0.65  3.02
reaction  0.12  0.91  1.25  2.40
gender    0.03  0.01 -0.02 -0.01  0.23
age       0.07 -0.04  0.74 -0.75  0.88 33.65

         cond  pmi   imprt rectn gendr age
cond      1.00
pmi       0.18  1.00
import    0.18  0.28  1.00
reaction  0.16  0.45  0.46  1.00
gender    0.13  0.02 -0.03 -0.01  1.00
age       0.03  0.00  0.07 -0.08  0.32  1.00
```

The mediation model



Compare regression to mediation (to correlation)





Moderation

1. Moderated multiple regression is merely the case of adding a product term
2. $y \sim x_1 * x_2$
3. which becomes $y \sim x_1 + x_2 + x_1 * x_2$
4. The product term will be highly correlated with the additive terms unless we zero center the data
5. All of this is done automatically in `mediate` or `lmCor` if we specify the moderator (and include the raw data)
6. Quadratic terms may also be specified.

Raw and Standardized Moderated regression

Using the `lmCor` or `mediation` function.

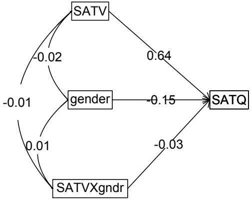
R code
```
lmCor(SATQ ~ SATV  * gender ,data=sat.act)
mediate(SATQ ~ SATV  * gender ,data=sat.act)
mediate(SATQ ~ SATV  * gender ,data=sat.act,std=TRUE)
```

```
Mediation/Moderation Analysis
Call: mediate(y = SATQ ~ SATV * gender,
data = sat.act)

The DV (Y) was  SATQ . The IV (X) was
  SATV    gender SATV*gender .
 DV =  SATQ
              slope   se     t       p
SATV           0.66 0.03 22.72 6.4e-86
gender       -37.05 6.85 -5.41 8.5e-08
SATV*gender   -0.07 0.06 -1.10 2.7e-01

With R2 =  0.44
R = 0.67 R2 = 0.44   F = 184.19 on 3 and 696 DF
p-value:  6.69e-88
```

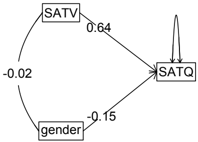
```
Mediation/Moderation Analysis
Call: mediate(y = SATQ ~ SATV * gender,
             data = sat.act, std = TRUE)

The DV (Y) was  SATQ . The IV (X) was
 SATV         gender SATV*gender .
 DV =  SATQ
              slope   se     t       p
SATV           0.64 0.03 22.72 6.4e-86
gender        -0.15 0.03 -5.41 8.5e-08
SATV*gender   -0.03 0.03 -1.10 2.7e-01

With R2 =  0.44
R = 0.67 R2 = 0.44   F = 184.19 on 3 and 69
  p-value:  6.69e-88
```

Compare moderated regression with normal regression

Moderation model



Regression Models



The correlation coefficient

1. Perhaps the most powerful and useful statistic ever developed
2. Special cases of the correlation are used throughout statistics.
3. The basic concepts of correlation are very straight forward
4. Many ways to be misled with correlations.

Aldrich, J. (1995). Correlations genuine and spurious in Pearson and Yule. *Statistical Science*, 10(4):364–376.

Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, 187(4175):398–404.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34(7):571–582.

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

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.

Pearson, K. P. (1910). *The grammar of science*. Adam and Charles Black, London, 3rd edition.

Preacher, K. J. and Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers*, 36(4):717–731.

Revelle, W. (2015a). Charles Spearman. In Cautin, R. L. and Lilienfeld, S. O., editors, *The Encyclopedia of Clinical Psychology*. John Wiley & Sons Inc.

Revelle, W. (2015b). Francis Galton. In Cautin, R. L. and Lilienfeld, S. O., editors, *The Encyclopedia of Clinical Psychology*. John Wiley & Sons, Inc.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 13(2):238–241.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, 83(2):213–217.

Waller, N. G. (2008). Fungible weights in multiple regression. *Psychometrika*, 73(4):691–703.

Wang, Z. and Rousseau, R. (2021). Covid-19, the yule-simpson paradox and research evaluation. *Scientometrics*, 126(4):3501–3511.

Wilks, S. S. (1938). Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika*, 3(1):23–40.

Yule, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika*, 2(2):121–134.