Psychology 360: Personality Research
Part II: Psychometric Theory
Part II: Multiple correlation and beyond

William Revelle

Department of Psychology
Northwestern University
Evanston, Illinois USA

October, 2022

Outline

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.

R code

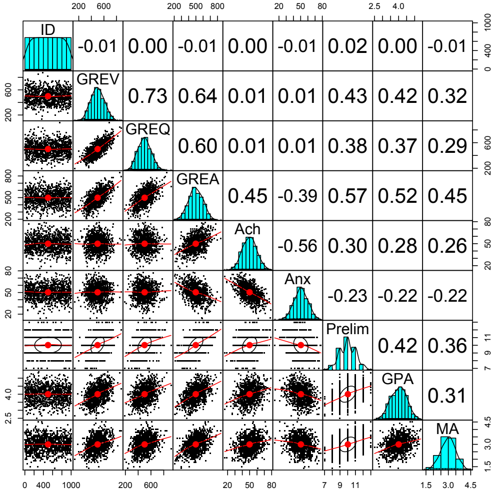
```
>  datafilename="http://personality-project.org/r/datasets/psychometri
>  mydata =read.table(datafilename,header=TRUE)  #read the data file
> describe(mydata,skew=FALSE)
```

	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='.').

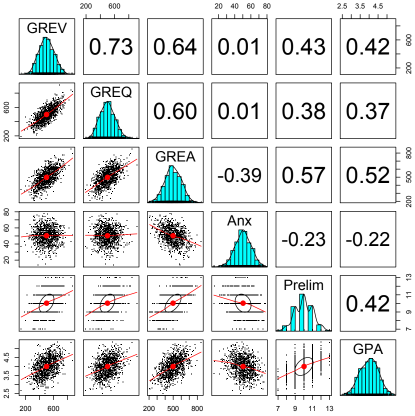
>pairs.panels(mydata,pch=".",gap=0) #pch='.' makes for a cleaner plot



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

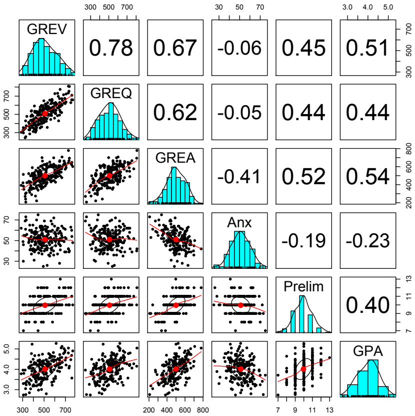
Use the pairs.panels function to show a splom plot. Select a subset of variables using the c() function.

>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 do interaction terms 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 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
```

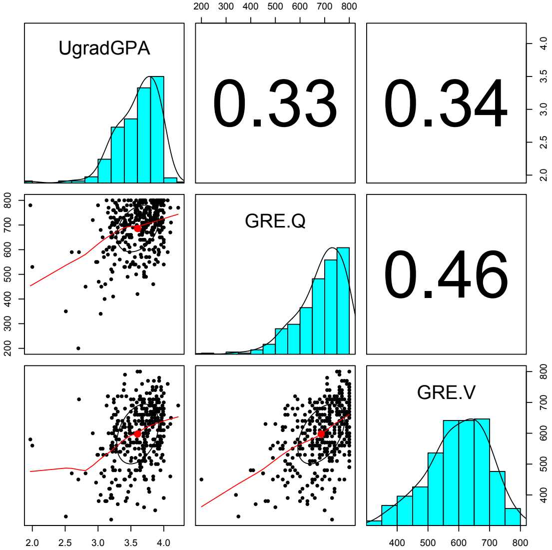
```
> lowerCor(z.data)

        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

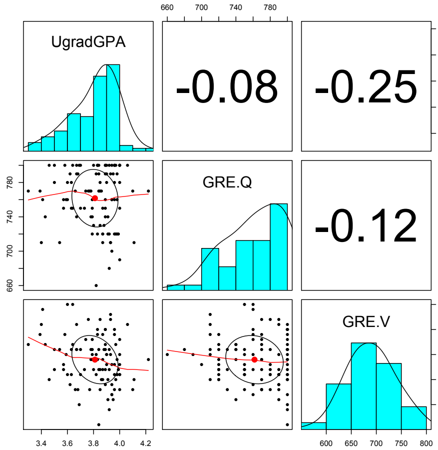
```
r <- lowerCor(my.data)  #find the original correlations
z <- lowerCor(z.data)   #find the z transformed correlations
lu <- lowerUpper(r,z,diff=TRUE) #combine into one matrix and take t

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

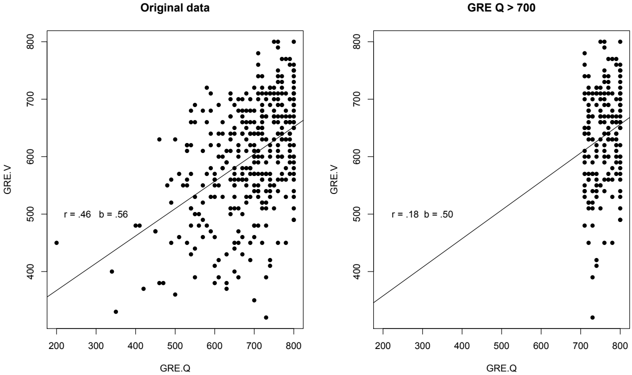


The effect of selection on the correlation



- 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

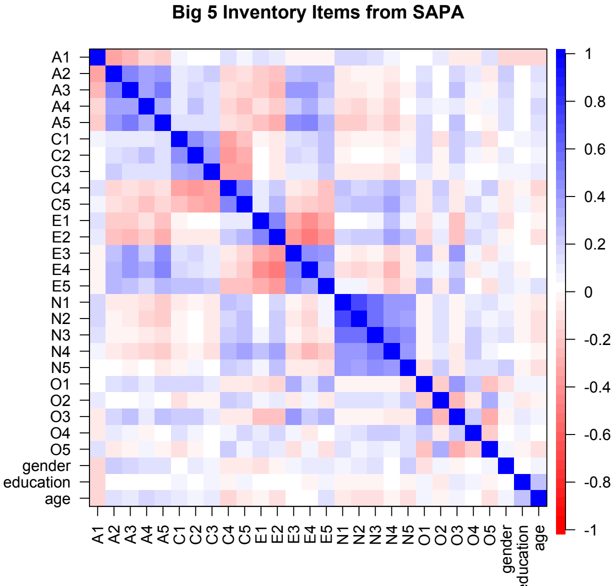


Although the correlation is very sensitive, regression slopes are relatively insensitive to restriction of range.

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 & 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

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

$$\phi = \frac{AD - BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}. \tag{1}$$

Correlation size $\neq$ causal importance

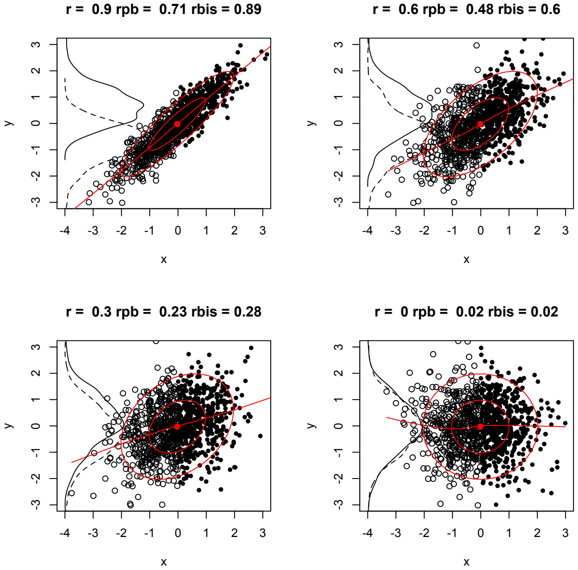
Table: The relationship between sex and pregnancy (hypothetical data)

	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		

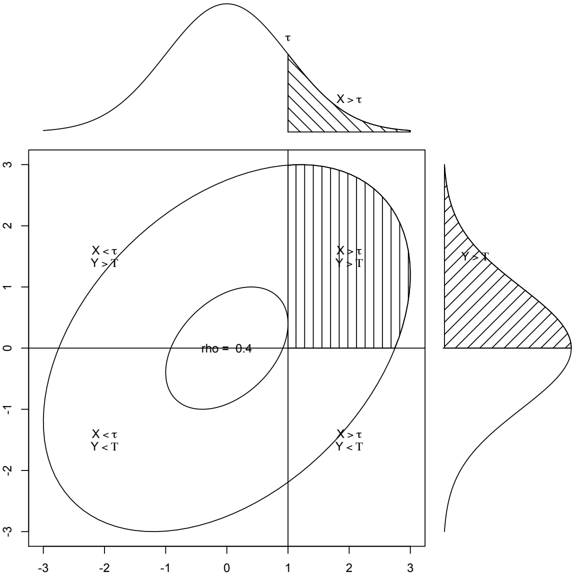
```
> sex <- c(2, 1041,0,6257)
> phi(sex)
```

```
[1] 0.04
```

The biserial correlation estimates the latent correlation



The tetrachoric correlation estimates the latent correlation



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
```

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

```
> 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. polychromic 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).

Cautions about correlations–The Anscombe data set

Consider the following 8 variables

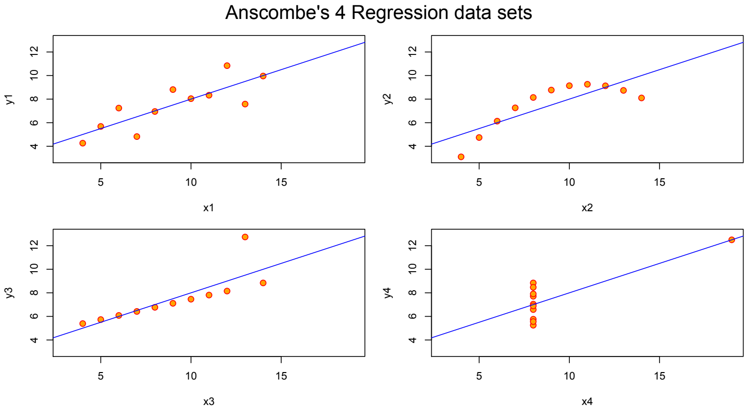
	var	n	mean	sd	median	trimmed	mad	min	max	range	skew	kur	se
x1	1	11	9.0	3.32	9.00	9.00	4.45	4.00	14.00	10.00	0.00	−1.20	1.00
x2	2	11	9.0	3.32	9.00	9.00	4.45	4.00	14.00	10.00	0.00	−1.20	1.00
x3	3	11	9.0	3.32	9.00	9.00	4.45	4.00	14.00	10.00	0.00	−1.20	1.00
x4	4	11	9.0	3.32	8.00	8.00	0.00	8.00	19.00	11.00	2.47	11.00	1.00
y1	5	11	7.5	2.03	7.58	7.49	1.82	4.26	10.84	6.58	−0.05	−0.53	0.61
y2	6	11	7.5	2.03	8.14	7.79	1.47	3.10	9.26	6.16	−0.98	0.85	0.61
y3	7	11	7.5	2.03	7.11	7.15	1.53	5.39	12.74	7.35	1.38	4.38	0.61
y4	8	11	7.5	2.03	7.04	7.20	1.90	5.25	12.50	7.25	1.12	3.15	0.61

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
```

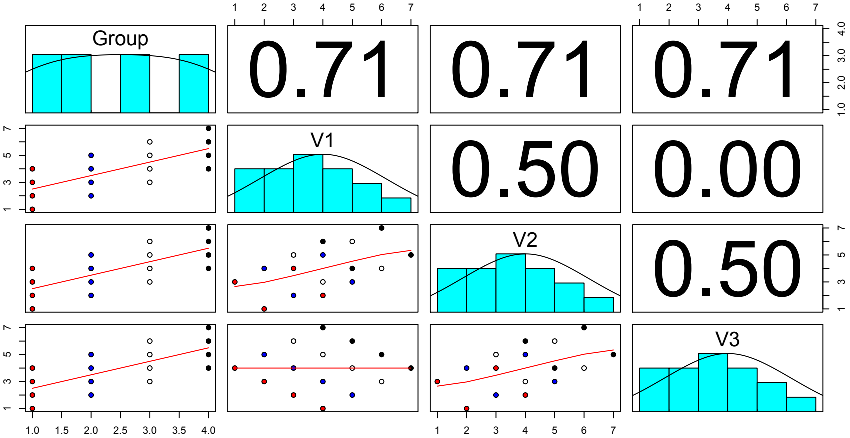
Cautions about correlations: Anscombe data set



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



Bias, or just Simpson's 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

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

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:

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

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

$$r_{xy} = Cov_{z_x z_y} = Cov_{\frac{x}{\sigma_x} \frac{y}{\sigma_y}} = \frac{Cov_{xy}}{\sigma_x \sigma_y} \tag{3}$$

Error of correlation

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?

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

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

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

$$V_r = V_y - r_{xy}^2 V_y = V_y(1 - r_{xy}^2) \quad (4)$$

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

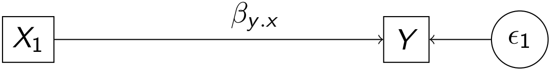
Variance and correlations

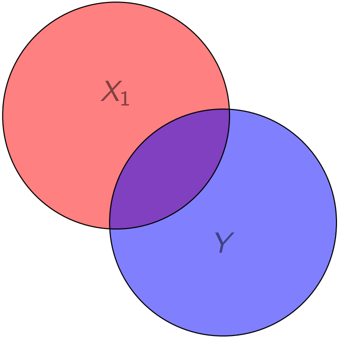
The various relationships between correlations, predicted scores, the variance of the predicted scores, and the variances of the residuals may be seen in the following table (10).

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

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

$$d_{xy} = \sqrt{2 * (1 - r_{xy})}. \tag{5}$$

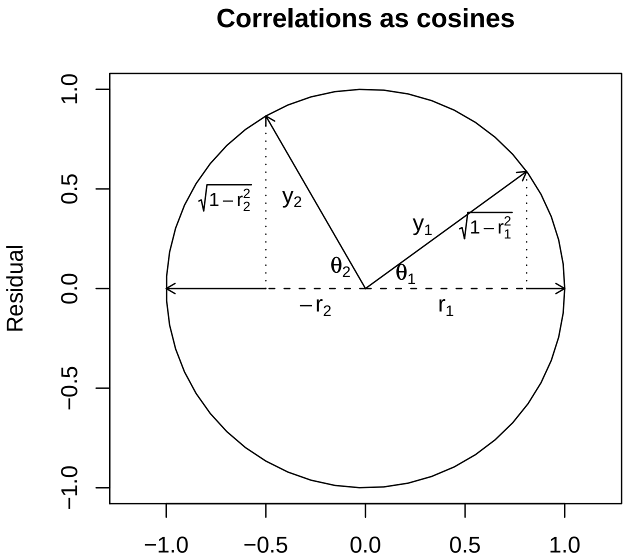
Distance, correlations, and the law of cosines

Compare this to the trigonometric *law of cosines*,

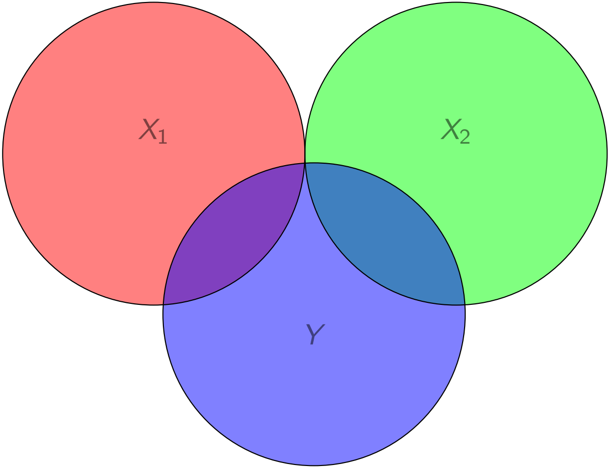
$$c^2 = a^2 + b^2 - 2ab \cdot cos(ab),$$

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

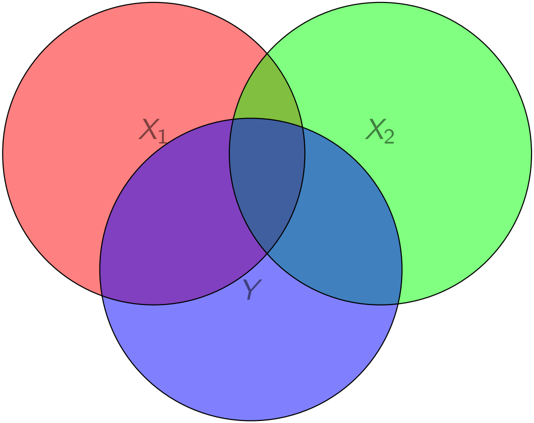


The Ideal model of predicting Y from X_1 and X_2



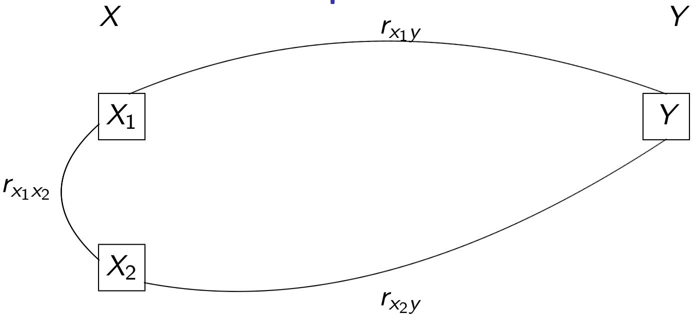
Variance in Y predicted by X_1 and X_2 if X_1 and X_2 are independent. $\hat{V}_y = V_y r^2_{x_1 y} + V_y r^2_{x_2 y}$

The usual case of predicting Y from X_1 and X_2

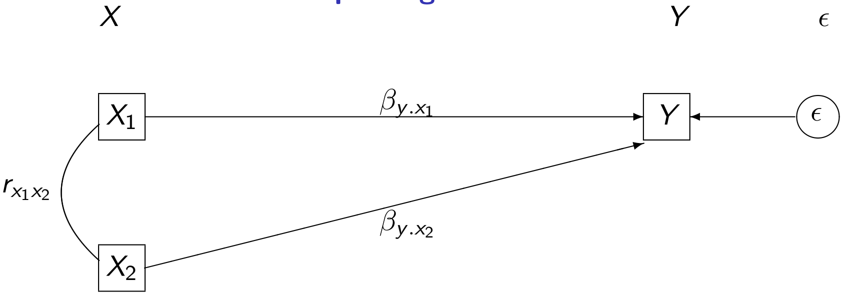


Variance in Y predicted by X_1 and X_2 if X_1 and X_2 - overlapping predictions $\hat{V}_y = V_y r^2_{x_1 y} + V_y r^2_{x_2 y} -$ overlap
But what is the overlap?

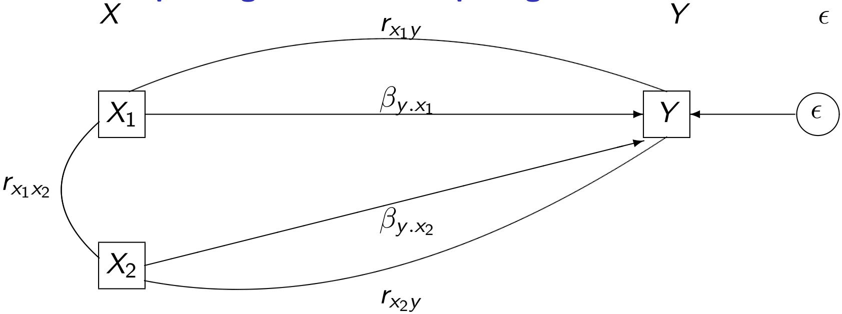
Multiple correlations



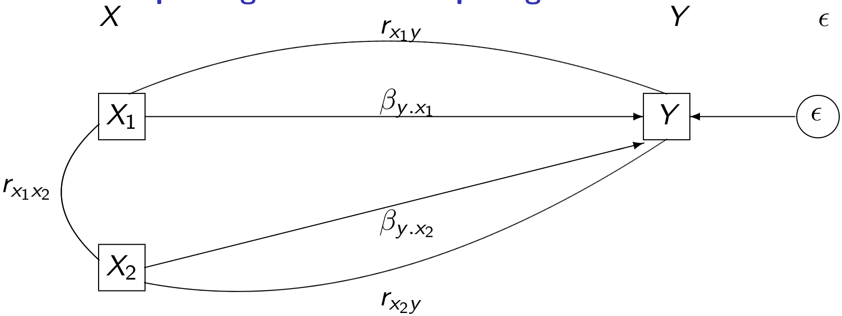
Multiple Regression



Multiple Regression: decomposing correlations



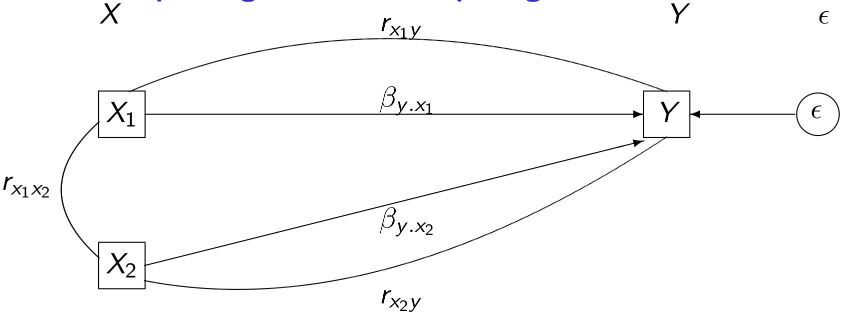
Multiple Regression: decomposing correlations



$$r_{x_1y} = \overbrace{\beta_{y.x_1}}^{\textit{direct}} + \overbrace{r_{x_1x_2}\beta_{y.x_2}}^{\textit{indirect}}$$

$$r_{x_2y} = \underbrace{\beta_{y.x_2}}_{\textit{direct}} + \underbrace{r_{x_1x_2}\beta_{y.x_1}}_{\textit{indirect}}$$

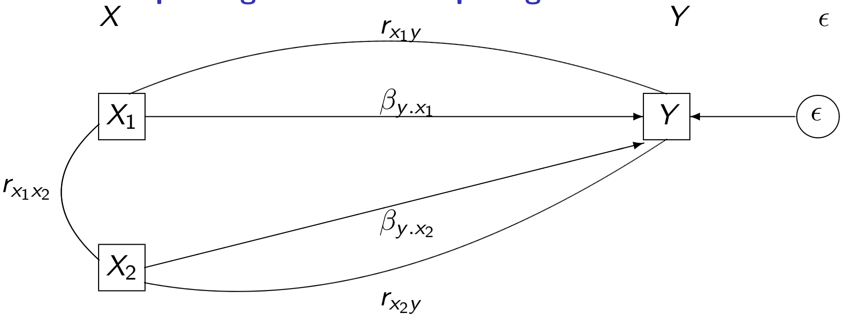
Multiple Regression: decomposing correlations



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

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

Multiple Regression: decomposing correlations



$$r_{x_1y} = \overbrace{\beta_{y.x_1}}^{direct} + \overbrace{r_{x_1x_2}\beta_{y.x_2}}^{indirect}$$

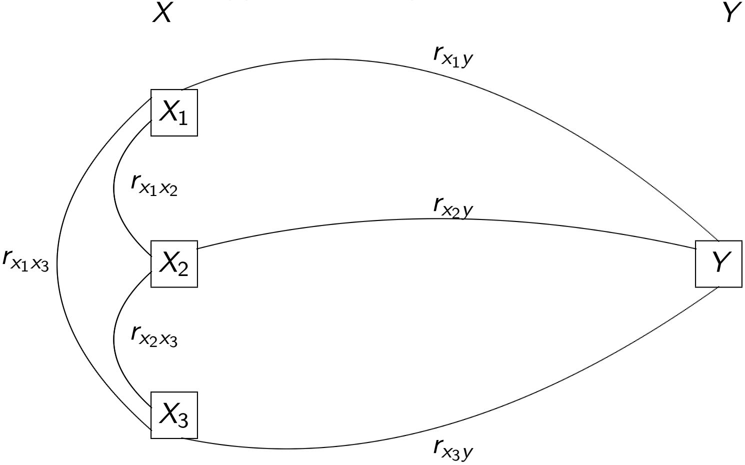
$$r_{x_2y} = \underbrace{\beta_{y.x_2}}_{direct} + \underbrace{r_{x_1x_2}\beta_{y.x_1}}_{indirect}$$

$$\beta_{y.x_1} = \frac{r_{x_1y} - r_{x_1x_2}r_{x_2y}}{1 - r^2_{x_1x_2}}$$

$$\beta_{y.x_2} = \frac{r_{x_2y} - r_{x_1x_2}r_{x_1y}}{1 - r^2_{x_1x_2}}$$

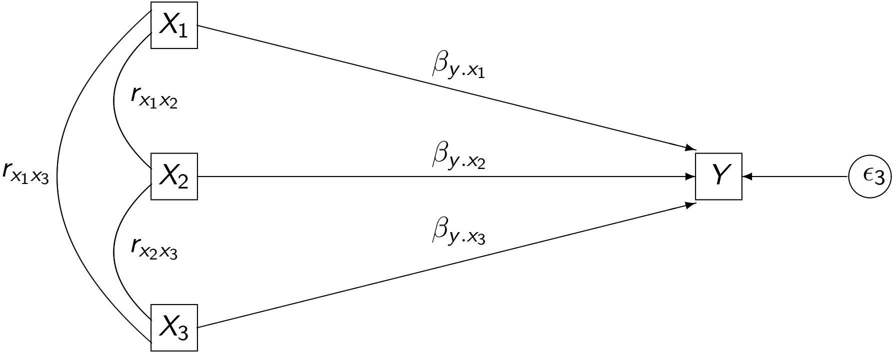
$$R^2 = r_{x_1y}\beta_{y.x_1} + r_{x_2y}\beta_{y.x_2}$$

What happens with 3 predictors? The correlations

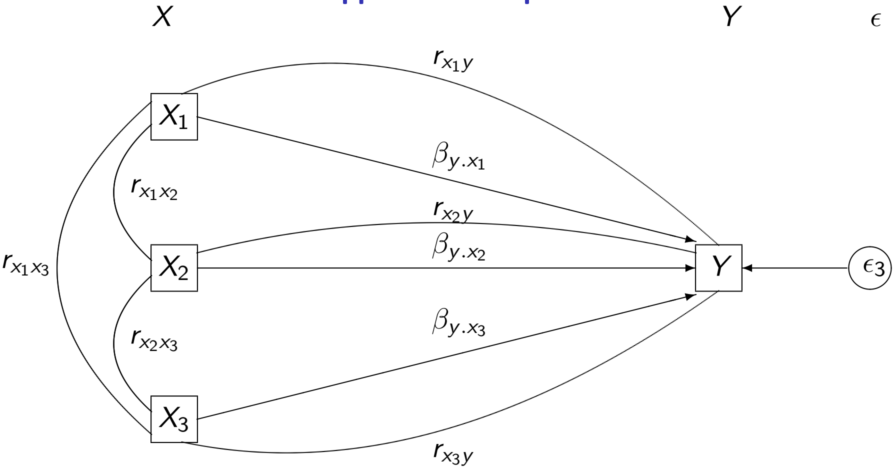


What happens with 3 predictors? β weights

X $\qquad$ Y $\qquad$ ϵ



What happens with 3 predictors?



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

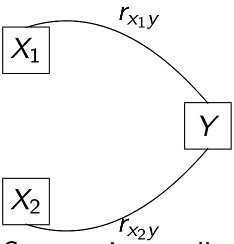
The math gets tedious

Multiple regression and linear algebra

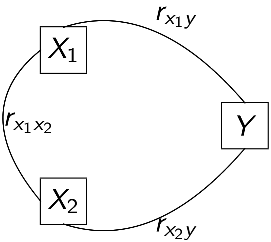
- Multiple regression requires solving multiple, simultaneous equations to estimate the direct and indirect effects.
    - Each equation is expressed as a $r_{x_i y}$ in terms of direct and indirect effects.
    - Direct effect is $\beta_{y.x_i}$
    - Indirect effect is $\sum_{j \neq i} beta_{y.x_j}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

3 special cases of regression

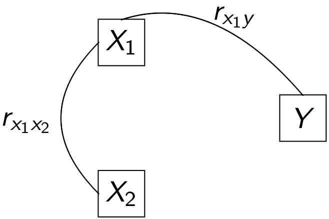
Orthogonal predictors



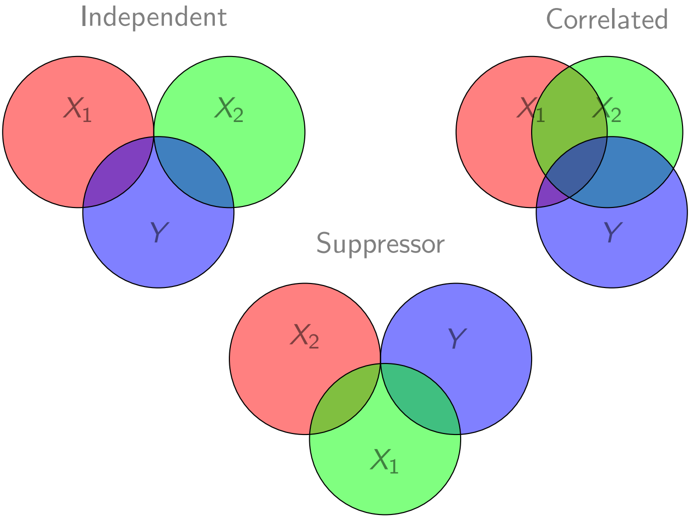
Correlated predictors



Suppressive predictors

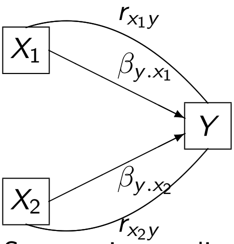


Three basic cases

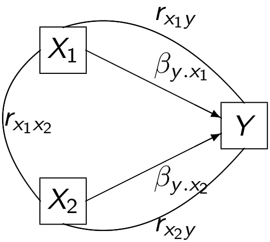


3 special cases of regression

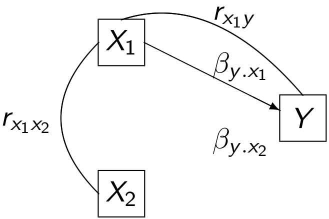
Orthogonal predictors



Correlated predictors



Suppressive predictors

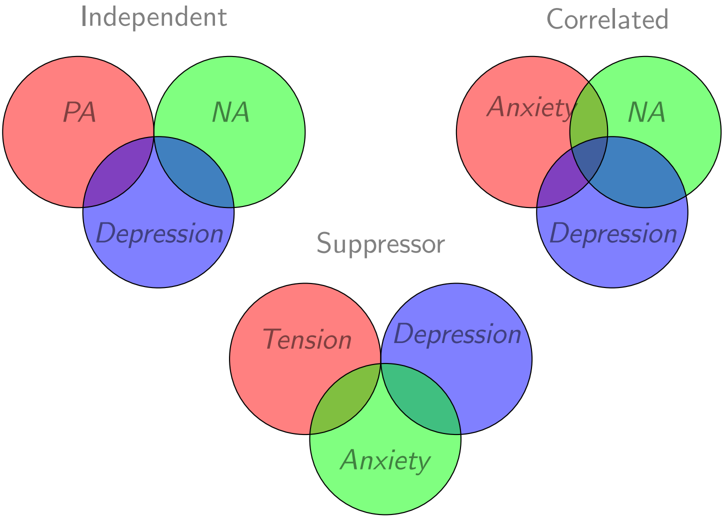


$$\beta_{y.x_1} = \frac{r_{x_1 y} - r_{x_1 x_2} r_{x_2 y}}{1 - r^2_{x_1 x_2}}$$

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

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

Three basic cases: Theoretical examples



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 Ã"***Ã• 0.001 Ã"**Ã• 0.01 Ã"*Ã• 0.05 Ã".Ã• 0.1 Ã

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

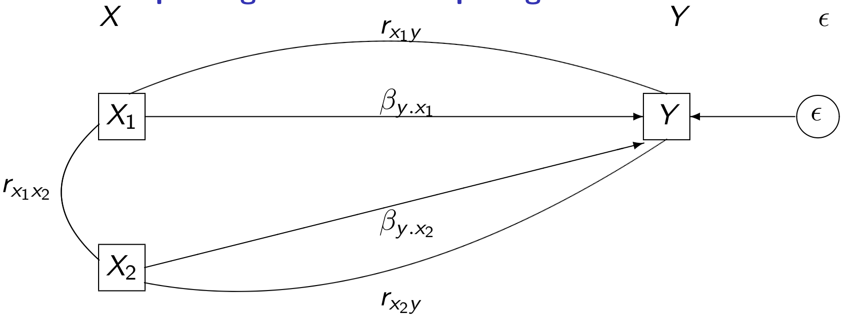
```
> mod2 <- lm(GPA~GREV, data=z.data)
> summary(mod2)

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 Ã”***Ã• 0.001 Ã”**Ã• 0.01 Ã”*Ã• 0.05 Ã”.Ã• 0.1 Ã

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 Ã”***Ã• 0.001 Ã”**Ã• 0.01 Ã”*Ã• 0.05 Ã”.Ã• 0.1 Ã

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



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

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

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

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

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

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 Ã”***Ã• 0.001 Ã”**Ã• 0.01 Ã”*Ã• 0.05 Ã”.Ã• 0.1 Ã

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 Ã”***Ã• 0.001 Ã”**Ã• 0.01 Ã”*Ã• 0.05 Ã”.Ã• 0.1 Ã

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=

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 Ã"***Ã• 0.001 Ã"**Ã• 0.01 Ã"*Ã• 0.05 Ã".Ã• 0.1 Ã

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

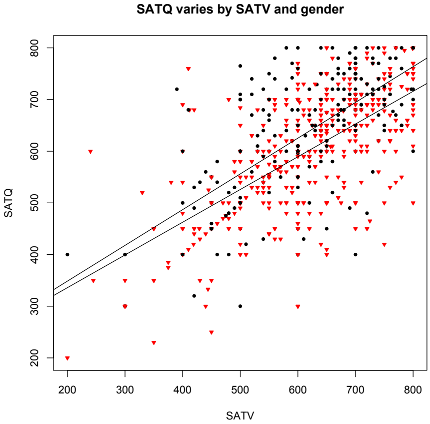
```
> mod4 <- lm(GPA ~ GREV + GREQ + GREA , data= cent[400:600,])
> summary(mod4)

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 Ã”***Ã• 0.001 Ã”**Ã• 0.01 Ã”*Ã• 0.05 Ã”.Ã• 0.1 Ã

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.act,scale=FALSE
>
summary(lm(SATQ~SATV * gender,data=c.sat))

Call:
lm(formula = SATQ ~ SATV * gender,  data = c.s

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

Coefficients:
                   Estimate  Std. Error  t  value  Pr(>
(Intercept)   -0.26696      3.31211
-0.081      0.936
SATV            0.65398      0.02926
22.350  < 2e-16  ***
gender        -36.71820      6.91495
-5.310  1.48e-07  ***
SATV:gender   -0.05835      0.06086
-0.959      0.338
---
Signif. codes:
0 Ã"***Ã• 0.001 Ã"**Ã• 0.01 Ã"*Ã• 0.05 Ã".Ã•

Residual standard error: 86.79 on 683 degrees
  (13 observations deleted due to missingness
Multiple R-squared: 0.4391      Adjusted R-sq
```

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
---
Signif. codes:  0 Ã"***Ã• 0.001 Ã"**Ã• 0.01 Ã"*Ã• 0.05 Ã".Ã• 0.1 Ã

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
```

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 m
          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
age         0.58      0.00 0.00 0.03    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`

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.

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